目　录

第一篇 绪 论

地质学与其他学科相比，研究工作的定量化程度是比较低的，形成这种状况的根本原因是由地质学的本身特点所决定的。地质学所研究的内容，几乎都是地球上久远以前发生的地质过程，目前留给人们可以观察或测量的地质现象都是经过长期地质演变后的遗留痕迹；也就是说，人们只能用不完全的残留信息去推断早已发生过的地质演变过程。所以，长期以来，地质学的主要研究内容就是对地质现象进行记实性描述、分类归纳、思维推理，最后形成地质认识或地质论断。目前的多数地质理论就是经过长期观察积累，其后又为大量事实所验证的所谓规律性理论；其中缺少验证或有待验证的一些论点，就是所谓的地质假说。众所周知，剖析近代的地质演变过程，以及"将今论古"的类比方法，对于不少地质理论的形成，都起到过积极的推动作用。

在地质学中引进定量研究方法，约始于上世纪40年代，但定量化的进程却非常缓慢。由于地质问题在地域上的差异与地质家思想方法上的差别，致使地质学界流派甚多，同一概念具多种定义者不胜枚举，同一问题有多种论断者也屡见不鲜。地质学中的这种状况必然促使一些人着手探索定量化的研究方法，以期对地质问题得出符合实际的统一认识。这就是数学地质学产生的历史背景。

第一章 数学地质的有关问题

从本世纪50年代末期开始，数学地质得到了迅速发展，主要与地质信息激增、找矿难度加大以及计算机技术的普及有关。

第一节 数学地质的现代含义

数学地质是地质学中出现较晚的一个分支，目前，数学地质的研究领域与研究内容还在不断地扩充与发展。多数的数学地质工作者认为：数学地质是地质学与数学互相渗透、紧密结合而产生的一门边缘学科。它是以数学为方法，以计算机为手段，对地质问题（包括地质理论问题及实际找矿问题）进行定量研究的实用性方法学，其最终目的是使地质学实现定量化研究。

石油数学地质是数学地质在石油地质中的应用，其主要任务是研究石油地质学中的有关定量问题，近年来它的主要研究内容是石油资源评价方法及其相应软件的开发。

数学地质的基本工作过程是，首先由地质家提出地质问题，并且建立相应的概念模型；通过分析研究，选用合适的数学方法，将定性的概念模型转化为定量的数学模型以及相应的

应用软件；最后对计算机运算得到的定量结果（包括数值与图形）进行地质解释，以期解决开始提出的地质问题。这一过程详见图1-1-1。

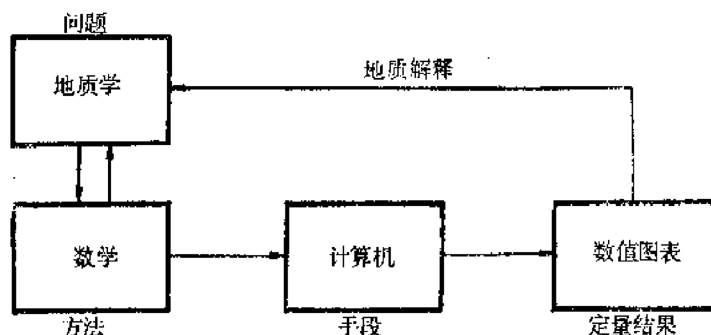


图1-1-1 数学地质的工作流程示意图

第二节 数学地质的主要内容

一、地质多元统计

地质多元统计是应用宏观统计方法研究地质问题的方法的统称。其中的多数方法是从数理统计学中直接移植过来的，少数方法是根据地质工作的实际需要，在移植的基础上逐步发展衍生出来的。

地质多元统计是数学地质的基础，也是石油数学地质的主要方法。本世纪70年代以前，数学地质的主要研究内容就是地质多元统计方法，及其在实际研究地质工作中的应用。因而，有人认为地质多元统计是数学地质发展过程中的第一台阶。

目前，地质多元统计方法已比较完善，它起到了使地质学向定量化方向发展的导向作用。但是，它在应用领域的广度上还远远不够，而且多数常规地质研究人员对地质多元统计方法也并不很熟悉。

地质多元统计方法包括回归分析、趋势分析、聚类分析、判别分析、因子分析、对应分析、典型相关分析、时间序列分析、非线性映射、马尔科夫链等主要数学方法。

近年来，在历次全国性的数学地质学术讨论会上，地质多元统计方面的学术论文数量最多，一般不少于论文总数的60%。这就说明地质多元统计仍然是数学地质的重要组成部分。稍作分析即可看出，地质多元统计之所以在长时间内能够得以存在和发展，这与地质学本身的特点有关。因为任何一个地质问题都是非常复杂的，即地质问题都具有时间久、空间广、因素多这三个基本特征。因而，地质人员总是试图借用统计分析方法从已知信息中获得一些规律，以便从定量角度分析地质问题。这也就必然使得地质多元统计在地质研究工作中具有广阔的应用领域。

二、矿产资源预测

自有人类以来，人们就开始了解他们的生存环境以求适应，开发矿业以图发展。地质学的重要意义在很大程度上与人类的找矿活动有关。因而，矿产资源预测一直是地质学的重要

研究内容。随着矿产资源的迅速开采，找矿难度越来越大，所以从本世纪70年代开始，矿产资源的定量预测就已成为数学地质的重要研究内容之一。实践证明，数学地质定量预测方法的有效性越来越明显，并已得到地质界的普遍重视。

无论是什么矿种，矿产资源预测的目的是要回答“有没有？有多少？”和“如果有，到哪找？”这两个基本地质问题。此外，对探区进行经济评价与勘探决策也应作为地质评价的延伸性工作。

矿产资源预测在石油部门称作石油资源评价，它是石油数学地质的主要组成部分，其主要内容包括：

（1）探区石油资源量的估算；

（2）确定探区中的有利勘探地带；

（3）石油勘探的经济分析。

目前，我国有关地质找矿部门根据各自的实际需要，针对所找矿种的地质特征，都在大力研究矿产资源的定量预测理论及其相应的计算方法与应用软件。例如，冶金部门近十年来一直大力发展“地质统计学”。地质统计学是南非金矿矿山工程师克里格（D.G.Krige）首先提出的预测金矿矿床的一种方法，后来经过法国的应用数学家马特隆（G.Matheron）加以理论化，而成为一个比较完整、自成体系的研究固体矿床，预测品位空间分布的专门方法。近年来，已在法国及法语系国家形成一支有影响的学派。由于这种方法是由克里格首先提出的，所以也称为“克里格法”。地质统计学研究的主要内容是区域化变量，并且通过建立变异函数，研究固体矿床空间变化、勘探方法与储量误差三者之间的定量关系。

三、地质数据库

地质数据库是计算机技术为地质人员服务的一个范例，也是地质学现代化的一个重要标志。地质数据库的出现，使地质人员可以从繁琐的收集资料等非研究性工作中解脱出来，它已成为科学管理地质信息的重要手段，从而为地质问题的定量化研究创造了有利条件。

数据库是存储在一起的相关数据集合，这些数据中无有害的或不必要的冗余，可为多种应用服务；数据的存储独立于使用它的程序，对于插入新数据，修改和检索原数据均能按照公用的和可控的方式进行。因而，一个完善的数据库应包括数据的存储、检索、更新、处理、显示、通讯、网络等多种功能。

数据库是60年代末出现的最新的数据管理技术，比较完善的数据库软件系统是70年代初完成的。地质数据库在美国、加拿大、法国、德国等西方国家发展很快，80年代以来，在许多国家已普及应用。目前，世界上大约已建成500多个大、中型地质数据库。这些数据库已涉及到地质学的各个分支领域，其中有些大型数据库已在一个国家甚至许多国家形成网络系统。比较著名的地质数据库有：

1．计算机化矿产资源信息库（CRIB）

这是美国地质调查所的矿产资源数据库。库内存有美国国内的4万多个矿床和矿产产地以及其他国家的6千多个矿床和矿产产地的有关记录。数据文件中包括：矿床位置，地质特征、储量、产量等多种数据。用户可以通过计算机网络系统在全世界500多个城市用电话查询和索取这个数据库中的地质数据。

2．北美石油数据系统（PDS）

这一数据库包括目前公开使用的10个石油地质数据库。库中存储了美国和加拿大的10万多个油气田的有关数据，文件内容包括：油气田的产量、生产井数、储量、圈闭类型、储集层时代、储集层厚度、油气性质、地层温度、地层压力、岩性等多种数据。

3. 井史控制系统（WHCS）

该数据库系统属于"石油信息公司"，这是世界上最大的一个油井数据库。数据文件中存有美国100多万口油气井的有关数据。

目前，虽然地质数据库的数据结构、各层次的命名术语等方面各有不同，但是，数据库的大的层次基本上可以归纳为4个级别，即子库、文件、项目、数据。

近年来，国内利用微型计算机管理地质数据的工作进展很快。但是，限于微型计算机的功能与存储空间都很有限，所以所建的数据库大体上相当于上面所说的4个级别中的文件级。因而，利用微型计算机管理地质数据，一般来说，只适用于某一专项地质工作。

四、地质过程的数学模拟

应用数学模拟方法研究地质历史演化过程，是探索地质基础理论的重要途径。近十年来，地质过程的数学模拟已成为数学地质的重要组成部分，其发展速度较快，例如，通过盆地模拟研究石油地质演变历史已成为一个热门课题。

由于地质过程的复杂性，所以数学模拟一般总是从简化模型入手，通常是从建立概念模型开始，最后转化为数学模型，再通过计算机的运算，从而得到对地质过程的定量描述。有时为了证实概念模型及其相应数学模型的可靠性，也需要进行物理模拟验证。

所谓地质模型可以理解为地质家对所研究地质问题的演化过程所持观点的形象化描述，简言之，地质模型就是对地质体系的一个表示或体现。其中，地质概念模型是指在对地质体系深刻理解的基础上，用定性方式表述地质体系演化过程及其量间关系的模型；而地质数学模型是指用定量方法描述地质体系演化过程及其量间关系的模型。概念模型是建立数学模型的基础，而由概念模型过渡到数学模型是对地质体系认识的深化。为了分析概念模型与数学模型的可靠性，经常采用实验手段对模型进行验证，例如用水槽试验模拟沉积过程；用泥巴试验、光弹试验模拟构造演化过程等等。对于这些实验一般称作物理模拟，也有人称作物理模型。

数学模型按其使用的数学方法又可分为确定型模型与随机型模型。然而，任何一个实际的地质过程都不可能是单一的确定型过程或随机型过程，而实质上都是这两种过程在时间、空间上的组合，因而一个完善的数学模型应该是包括这两种过程的复合模型。目前，只使用单一的确定型模型或随机型模型来研究地质历史演化过程，都是限于人们认识上的不完全或者是把复杂的地质问题进行简化所造成的。

以上四个方面就是现阶段数学地质研究的主要内容。这四个方面相对独立又相互关联，虽然每个方面的研究内容各有侧重，但是它们都是为了一个统一的目的，那就是使地质学加快定量化的进程，并且最终实现地质学的定量化研究。

一个学科实现数学化（或叫定量化）是这门学科成熟的标志之一。当今是科学技术飞速发展的时代，任何一个学科都在汲取数学上的成熟方法以及最新进展，或者根据本学科的需要向数学界提出新问题，从而促进数学的发展并服务于本学科。地质学的数学化就是用数学语言描述地质学中的定义、概念、规律等等，从而使地质学由目前的以定性描述为主转变到全面

的定量描述。

地质学实现定量化十分困难。除了地质问题本身十分复杂外，还有其他许多难点，例如同一地质概念的多种含义问题，观测手段的精度问题，地质数据的代表性问题等等都在很大程度上阻碍着地质学实现定量化。沉积学上“相”的概念，有人统计竟多达几十种含义，类似情况在地质学中屡见不鲜。出现这类情况，多半是由于地质学家工作经历的局限性，而使他们所提出的概念带有浓郁的地域性特色。由于上述种种原因致使地质学的定量化进展十分缓慢，定量化程度大大地落后于其他学科。这种情况的长期延续，以致在地质学界的不少人中产生一种偏见，即认为地质学不需要定量化，也不可能定量化。这种偏见也在很大程度上阻碍着地质学的定量化进程。

鉴于此，地质学的定量化将需要一个相当长的历史阶段，预计需要几代人的不懈努力才能逐步实现。

第三节　数学地质的历史简述

数学地质的发展历史大体上可分为如下三个阶段：

一、发起阶段（1840~1945）

这个阶段的特点是少数人在个别的地质研究中引进了数学方法，工作量很少，成果分散。

这一阶段的重要事件大都是一些代表人物的个人成就。如，1840年英国的莱伊尔（Lyell）通过古生物化石的统计分析，进行了第三纪地层的划分；1914~1934年间，俄国的列文生—列兴格（Левинсон—Лесинг）通过研究火成岩岩浆系数的频率分布，进行了流纹岩、安山岩、玄武岩的分类；美国的克鲁宾（W.C.Krumbein）从1934年开始进行的沉积作用和地层的统计分析工作很有影响，因而成为美国数学地质的奠基人；苏联的维斯捷列乌斯（А.Б.Вистелус）于1944年发表了“分析地层学”一文，比较完整地提出了用定量方法研究地质学的初步思想。他多年从事数学地质方面的研究工作，是苏联数学地质的创始人，曾任国际数学地质协会的第一任主席。

20世纪40年代，美国、德国开始研制电子计算机，为数学地质尔后的发展作了技术准备。可以说，数学地质的发展与计算机技术的发展是密切相关的。

二、形成阶段（1945~1968）

这一阶段的特点是数学地质方法已在地质学的各个分支领域得到普遍的应用，单变量及多变量的统计方法已日趋完善，数学地质已成为地质学中的一个独立的分支学科。电子计算机已成为数学地质的重要技术手段。

1946年美国的宾夕法尼亚大学研制出了ENIAC电子计算机，1952年研制出数字绘图仪，1953年研制了第一个FORTRAN语言编译系统。1954年美国已成批生产IBM650计算机。1958年开始生产第二代电子计算机，同年克鲁宾首次在地质杂志上公布了应用于地质方面的计算机程序。从1961年开始，美国的亚利桑那大学召开了一系列电子计算机在矿产工业中应用的学术讨论会。1963年第三代电子计算机试制成功。由于电子计算机在地质学上已得

到广泛应用，而导致数学地质文献的数量激增，1963年全世界数学地质方面的文献已超过了100篇/年。1967年美国石油地质工作者协会成立了电子计算机存储和索取委员会，同年，国际地质科学联合会成立了地质数据存储、自动处理和索取委员会（COGEODATA）。

1968年在法国巴黎召开的国际地质会议上成立了国际数学地质协会（IAMG），并且开始出版国际数学地质协会志。至此，数学地质在地质界已被确认，而成为一个独立的分支学科。

三、发展阶段（1968～现在）

这一阶段的特点是数学地质向更高的水平发展，愈来愈多的数学方法被应用到地质学的研究工作中，并且逐渐形成了地质多元统计、矿产资源预测、地质数据库、地质过程数学模拟四个主要发展方向。

我国的数学地质工作起步于本世纪50年代，当时有少数人在地质多元统计方面作过一些研究工作。例如赵鹏大、徐道一等人在地质多元统计应用方面以及介绍国外数学地质的有关情况方面曾起到先导作用。

我国的数学地质研究工作是本世纪70年代初才开始大面积发展起来的。当时，中国科学院、地质矿产部、武汉地质学院、长春地质学院、成都地质学院、云南大学、北京大学、石油工业部、冶金工业部、煤炭工业部、核工业部等研究单位和部门都已相继开展了数学地质研究工作，从1978年开始大约每隔两年召开一次全国性的数学地质学术讨论会。

1978年10月在杭州由中国科学院地质研究所与地质矿产部地科院矿床所联合召开了“第一届全国数学地质学术讨论会”。

1981年4月在长沙召开了“第二届全国数学地质学术讨论会”，会议期间成立了中国地质学会数学地质专业委员会。

1983年4月在四川乐山由中国地质学会数学地质专业委员会与中国石油学会石油地质专业委员会联合举办了“数学地质在石油资源评价及地质勘探中的应用学术讨论会”。

1985年4月在广州由中国地质学会数学地质专业委员会与煤炭学会煤田地质专业委员会联合举办了“数学地质在煤田地质中的应用学术讨论会”。

1986年11月在湖北宜昌由中国地质学会数学地质专业委员会召开了“第三届全国数学地质学术会议”。

1987年10月在山东泰安由中国地质学会数学地质专业委员会召开了“全国青年数学地质工作者学术讨论会”。

1988年10月在河北石家庄由中国地质学会数学地质专业委员会与中国核学会铀矿地质专业委员会联合召开了“全国铀金等矿产资源评价学术讨论会”。

1990年4月在成都由中国地质学会数学地质专业委员会召开了“第四届全国数学地质学术会议”。

此外，各有关部门根据工作需要也曾召开过多次中、小规模的数学地质学术讨论会以及各种类型的数学地质学习班。

上述各种会议和学习班大大地促进了我国数学地质学的发展。目前，全国从事数学地质研究的专业人员估计有2000人，而使用数学地质方法进行地质研究的人员数量就更多了。这些人员是促进我国地质研究工作向定量化方向发展的主要力量。

第四节　数学地质的展望

为了加速地质学向定量化研究的进程，数学地质学科在今后的发展过程中，需要着重注意如下几个问题：

1．从地质学最基础的概念开始定量化

目前，数学地质的全部研究内容几乎都是在接受或顺应现有地质概念的基础上进行的，所以，实际上其研究成果只能是现有定性地质概念的量化补充与延伸；或者说，数学地质的研究成果只是传统研究结果的佐证。如果按着这种方式发展下去，数学地质学的发展很难打破地质学的传统观念。

为此，需要认真分析一下现有多数地质概念的构成。事实上任何一个地质概念都有其复杂的内涵，直接用简单的定量方法显然无法表示这样的复杂内涵.因此，需要把复杂的地质概念分解为可用定量方法表示的单一概念，然后再把这些简单的定量概念合成为复合的定量地质概念。也就是说，从地质学中最基础的概念开始定量化，是使地质学走向定量化研究的出发点。

2．建立能够确切表示地质过程的数学模型

目前已有的单一确定型或随机型数学模型，都不足以表示复杂的地质历史过程，这也是数学地质目前尚不能解决地质学定量研究的根本症结。如何在时间上、空间上配置出由确定型与随机型两种模型表示的复合模型，即建立能够确切表示地质过程的数学模型是数学地质学科的核心问题。

此外，数学地质工作者在充分利用现有已用于地质研究的数学方法基础上，还要根据地质学实际研究工作的需要，提出新的数学问题，形成专门解决各类地质问题的数学方法。

3．地质过程的数值模拟应从简单的地质体着手

目前，国内外已有不少的研究人员着手盆地模拟研究工作。但是，一个沉积盆地的发展历史是相当复杂的，从寻找油气资源的角度来看，沉积盆地的演变历史几乎包括了全部石油地质学的内容。如果我们对沉积盆地中比较单一的地质体，例如三角洲、浊积扇、河道砂等尚且不能深入剖析，又如何能够说清由这些单一地质体在时间场与空间场中搭配构成的沉积盆地呢？因此，研究单一地质体的数值模拟工作应作为地质过程数值模拟的基础工作。

4．通过物理模拟为建立数学模型提供依据

目前的许多地质概念中往往包含一些经验性的成分，因而由这些概念建立的数学模型就未必可靠。一种可行的途径就是通过实验室的物理模拟去验证地质家的概念。但是，物理模拟无法克服地质过程的时间效应，它只是地质过程的一个近似。

5．加强地质信息的管理与利用

地质信息应当包括地质数据、地质图形、地质知识，因此，地质数据库应包括数据库、图形库、知识库。而地质信息的利用应当包括信息分解、信息提纯、信息综合。地质信息是地质研究工作的唯一依据，而加强地质信息的管理与利用是实现地质学定量研究的基本条件。

6. 加强地质学与数学以及其他相关学科的渗透

地质学实现定量化研究将需要地质专业人员与数学专业人员以及其他相关学科专业人员之间的密切配合。这些相关学科包括物理学、化学、计算机技术、实验室技术等。对地质学实现定量化研究，学科间的渗透、交流是十分重要的。

第二章 地质信息及其分类

地质信息是进行数学地质研究的基本依据，地质信息可分为地质数据、地质图件及地质观点三大类。

第一节 地 质 数 据

地质数据是指由数值或代码描述地质特征的数据，可分为观测数据、复合数据及经验数据三种类型。

一、观 测 数 据

观测数据是由各种观测手段直接测量得到的数据。由于没有经过任何加工处理，所以也称为原始数据。

观测数据按其自身的特点和运算规则又可分为定性数据和定量数据。定性数据包括名义型数据和有序型数据，定量数据包括间隔型数据和比例型数据。

1. 定性数据

定性数据是指不能用数值描述，而只能用符号或代码描述的观测数据。

（1）名义型数据 名义型数据没有明确的数量概念，并且数据之间也没有次序关系，只能用符号或代码形式表示。例如，描述岩石颜色的红、绿、灰、黑等等。又如，假若不深究岩性的详细概念，那么，砂岩、泥岩、石灰岩等岩石类型也可以认为是名义型数据。

名义型数据的量间关系只存在“相等”（＝）或“不相等”（≠）的关系。例如，红色＝红色；砂岩≠石灰岩。

（2）有序型数据 有序型数据之间没有明确的数量概念，但是数据之间有次序关系，常以等级符号或等级代码形式表示。例如，鉴定岩石相对硬度的摩氏标准，硬度由小到大分为10个级别，即：滑石、石膏、方解石、萤石、磷灰石、长石、石英、黄玉、刚玉、金刚石；又如，沉积盆地按面积大小可分为大型盆地、中型盆地、小型盆地；再如，生油岩中的干酪根类型可以分为Ⅰ型、Ⅱ型、Ⅲ型等等。

有序型数据之间除有相等、不相等关系外，还有“大于”（＞）及“小于”（＜）的关系。例如，滑石硬度＜金刚石硬度；从生油潜力上看，Ⅰ型干酪根＞Ⅲ型干酪根。

2. 定量数据

定量数据是指能用数值大小来描述的观测数据。

（1）间隔型数据 间隔型数据有明确的数量概念，可以用数值形式表示。例如，以基准海平面起算的地层分层数据就是典型的间隔型数据。

间隔型数据之间除了有相等、不相等与大于、小于关系外，还可以定量地说明x_i与x_j相互之间大多少或小多少，即任何两个间隔型数据x_i与x_j之间的差（x_i-x_j）是有实际意义的。

例如，两个地层分层数据的观测值分别为$x_i=+50m$，$x_j=-250m$，

$$x_i-x_j=50-(-250)=300m$$

若x_i与x_j分别表示上第三系的顶界与底界，则（x_i-x_j）为上第三系的厚度，即上第三系的厚度为300m。

（2）比例型数据　比例型数据有明确的数量概念，可以用数值表示。任何两个比例型数据x_i与x_j，不仅差（x_i-x_j）有实际意义，而且比值x_i/x_j也有实际意义，即可以说x_i比x_j大（或者小）多少倍。

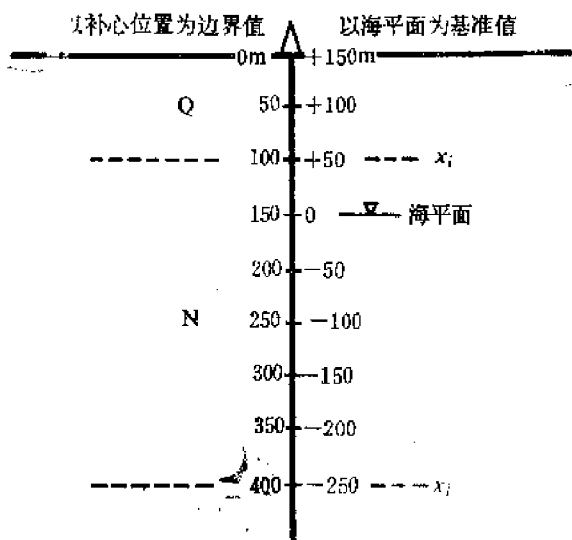


图1-2-1　比例型数据与间隔型数据的差别

比例型数据与间隔型数据是不相同的两种数据类型，二者的区别为比例型数据是以0为边界值的定量数据，即比例型数据是由正实数组成的数据集合，其中的最小值为0，而间隔型数据是由实数组成的数据集合，其中0不是边界值。

为了说明比例型数据与间隔型数之间的差别，现举例如下，见图1-2-1。

以补心位置作为0值起算的地层分层数据就是比例型数据。如果x_i与x_j分别表示晚第三纪地层的顶界与底界，那么

$$x_i-x_j=100-400=-300m$$

地层厚度不能为负值，所以$|x_i-x_j|=300m$。即第三纪地层的厚度为300m。而

$$x_i/x_j=100/400=\frac{1}{4}$$

就是说顶面x_i的深度为底面x_j的深度的四分之一。或者说，x_j的深度为x_i深度的四倍。

而以海平面为基准值的地层分层数据则是间隔型数据。如果x_i与x_j分别表示晚第三纪地层的顶界与底界，那么

$$x_i-x_j=50-(-250)=300m$$

即第三纪地层的厚度为300m。但是

$$x_i/x_j=50/-250=-\frac{1}{5}$$

则无任何实际地质意义。

此外，以摄氏度（℃）计量的地层温度就是间隔型数据；而以绝对温度（K）计量的地层温度则是比例型数据。

需要指出，多数的定量地质数据都是比例型数据，因为这些地质数据都只取正值，而且以0作为最小值。

二、复合数据

复合数据是指由定量观测数据（或者经过定量化处理后的定性数据），经过有限次数学运算后得到的有实际意义的综合性数据。

随机变量的特征值都是复合数据，例如平均值$\bar{x}$、子样的标准差σ

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\sigma=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

此外，极差、众数等也都是复合数据。

又如，熵也是一个典型的复合数据

$$H(x)=-\sum_{i=1}^{N}p_i\log P_i$$

古生物种属的组合常常作为研究古代沉积环境的重要指标，而复合数据熵$H(x)$则能较好地反映古生物群落的种属组合情况。

三、经验数据

经验数据是指在大量研究了地质现象和地质规律后，经过归纳（包括数学处理）而得到的经验值。经验数据的地质含义是明确的，一般情况下它是大量地质信息的综合反映。但是，经验数据究竟受哪些地质因素影响，以什么方式影响，即经验数据与影响它的地质因素之间的数学表达式是什么，暂时还不清楚。

石油资源评价工作中经常使用经验数据，如单储系数、聚集系数、排烃系数、单位产率等等都是经验数据。

由于每位地质家工作经历的局限性，所以，经验数据往往具有明显的地域性特征。因而在引用经验数据时要特别注意对比地质条件上的相似性。不加选择地引用将会得出错误的结果。

第二节　地质图件与地质观点

地质图件包括地质图形及地质变量的曲线。地质图件一般以线条、符号、颜色、灰度、等值线等表示，并且附以相应的数据和文字说明。

绝大多数地质问题仅用地质数据描述是不够的，而经常要用地质图件表示。地质图件中含有丰富的地质信息，可以形象地反映出地质信息之间的相互关系。地质图件一般可以分为观测图件和观点图件。

一、观测图件

观测图件是指由地质人员通过目测或仪器测量地质现象后生成的地质图件。这种图件一般能够真实、客观地反映地质现象，因而观测图件可看作是地质体系的观测模型。下面各类图件都是观测图件。

1. 地质图

地质图包括各种不同比例尺的地质图、剖面图、柱状图等图件。

2. 物探图

物探图包括由重力、磁法、电法、地震等各种地球物理勘探手段得到的记录所生成的图件。

3. 化探图

化探图包括由地球化学勘探得到的烃类及各种微量元素、化合物等化验分析数据生成的等值线图。此外，由放射性勘探得到的图件也经常与化探图件同时使用。

4. 遥感遥测图

遥感遥测图包括人造地球资源卫星各波段的扫描图片及航空侧视雷达图片。

二、观点图件

观点图件是指由观测数据、观测图件经过地质家加工后形成的图件。这种图件一般是以明确反映地质家主观观点为其特征。因而观点图件可看作是地质体系的观点模型。

由于地质问题的复杂性，而使许多的地质现象和地质过程的形成、演化等都并不完全清楚，因而地质家们常以丰富的想象力进行思维推理，由观测到的地质资料经过主观加工生成不同风格、不同流派的观点图件。例如大地构造图、岩相古地理图、古地貌图、矿产资源综合评价图都是典型的观点图件。

三、地质观点

地质观点是地质学家根据自己的工作实践，归纳总结出来的有一定适用范围的地质理论或地质假说。

地质观点多数是以文字叙述方式保存于各种地质文献中或地质家的头脑中。无疑，地质观点是地质学的重要财富。特别是今后随着计算机人工智能技术的发展，收集和整理地质观点是非常重要的。通常，地质观点可按生油、沉积、构造等地质分支学科进行分类整理以备使用。

不可否认，地质观点与经验数据一样，有其局限性和片面性。所以，在运用地质观点时要特别注意对比地质条件方面的相似性。

第三章　地质数据的预处理

地质数据的预处理往往是地质问题定量计算过程中不可缺少的一个重要环节。需要进行预处理的原因是多种多样的，但是，最根本的原因是定量计算过程中要求数据是规范化的。因而，研究地质数据的预处理已成为数学地质学的重要内容之一。

第一节　定量数据的标准化

地质学家经常使用地质数据表来描述地质问题，以表现地质问题的多因素特点。例如，在一个沉积盆地中已发现了n个地质圈闭，每个圈闭可以看作是一个地质样品。为了表示每个地质圈闭的特征，可在数据表中列出m项地质指标（也称地质变量），如果是构造型地质圈闭，则圈闭面积、闭合度、长短轴比、埋藏深度、断层条数等等都可以作为描述圈闭的地质变量。

这种数据表即为数据矩阵，它是由（$n\times m$）个数据组成，每个数据称为矩阵元素。如果把某个样品的m项变量排成一行，则构成如下n行m列的数据矩阵：

$$X=[x_{ij}]_{n\times m}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$$

数据矩阵中的i行表示第i个样品的m项变量的观测值；j列表示第j项地质变量的n个样品的观测值。矩阵中x_{ij}的第一个脚标i表示样品的编号，第二个脚标j表示变量的编号，即x_{ij}表示第i个样品的第j项地质变量。

使用定量数据时，由于各个地质变量的单位、量纲往往是不相同的，而且各个变量的数值大小、变化范围也是不相同的，因此，直接用原始数据进行计算是不合适的，显而易见，如果对原始数据不进行处理，在计算结果中就会突出那些数值大的地质变量的作用，而降低那些数值小的地质变量的作用。对定量数据进行标准化的目的就是为了克服这些不合理的因素，使单位不同、量纲不同、大小不同、变程不同的各个地质变量，通过变换而成为某种规范尺度下的地质变量。

例如，某个探区已经发现了5个地质圈闭，为了描述这些圈闭的地质特征，选用了圈闭面积、闭合高度、长短轴比、埋藏深度4项地质变量，见表1-3-1。

根据所研究问题的实际需要，定量数据的的标准化可分为对变量的标准化变换与对样品的标准化变换。但在实际研究工作中，多数情况下是要求把各种变量变换为同一尺度下的规范化变量；少数情况下也需要对样品进行标准化变换。

表1-3-1 地质圈闭数据表

圈闭编号 地质变量	1	2	3	4	5
圈闭面积(10^2m^2)	1000	250	100	10	40
闭合高度(m)	500	150	70	200	100
长短轴比	1.5	1.0	3	2	5
埋藏深度(m)	2000	2200	1500	1800	2500

表1-3-1是一个四行五列的矩阵，即：

$$X=[x_{ij}]_{4\times5}=\begin{pmatrix}1000 & 250 & 100 & 10 & 40\\ 500 & 150 & 70 & 200 & 100\\ 1.5 & 1.0 & 3 & 2 & 5\\ 2000 & 2200 & 1500 & 1800 & 2500\end{pmatrix}$$

这是一个行为变量，列为样品的矩阵。矩阵中第i行第j列上的元素为x_{ij}，这里令标准化变换后的元素为x'_{ij}。

一、总和标准化

对变量进行总和标准化变换时，是将变量的各样品观测值都变换为它与该项变量所有样品观测值总和的比值。因而，变换后的矩阵元素都变换为（0,1）开区间中的小数，而且变换后的矩阵X'中，每个变量的所有样品观测值之和等于1。其变换公式为

$$x'_{ij}=\frac{x_{ij}}{x_{i\cdot}} \qquad (i=1,2\cdots,m;\ \ j=1,2,\cdots,n) \qquad (1\text{-}3\text{-}1)$$

其中 $$x_{i\cdot}=\sum_{j=1}^{n}x_{ij} \quad (i=1,2,\cdots,m) \qquad (1\text{-}3\text{-}2)$$

对表1-3-1的矩阵进行变换时，首先要计算矩阵的各个行和$x_{i\cdot}$，按（1-3-1）及（1-3-2）式变换后得到X'，而$x'_{i\cdot}=1$。

$$X=\begin{pmatrix}1000 & 250 & 100 & 10 & 40\\ 500 & 150 & 70 & 200 & 100\\ 1.5 & 1 & 3 & 2 & 5\\ 2000 & 2200 & 1500 & 1800 & 2500\end{pmatrix} \quad \begin{matrix}x_{1\cdot}=1400\\ x_{2\cdot}=1020\\ x_{3\cdot}=12.5\\ x_{4\cdot}=10000\end{matrix}$$

$$X'=\begin{pmatrix}0.714 & 0.179 & 0.071 & 0.007 & 0.029\\ 0.[illegible]49 & 0.147 & 0.069 & 0.196 & 0.098\\ 0.120 & 0.080 & 0.240 & 0.160 & 0.400\\ 0.200 & 0.220 & 0.150 & 0.180 & 0.250\end{pmatrix} \quad \begin{matrix}x'_{1\cdot}=1\\ x'_{2\cdot}=1\\ x'_{3\cdot}=1\\ x'_{4\cdot}=1\end{matrix}$$

在二维情况下，对变量进行总和标准化变换时，在空间关系上是将原变量点径向投影到单位弦L上，径向是指由坐标原点到变量点的半径方向。这里约定，在下面的变换图形上用空心小圈表示原变量点，用实心黑点表示变换后的变量点。

例如，有两个样品、5个变量组成的数据矩阵，其中行为变量，列为样品。变换后的矩阵X'中，每个变量的两个样品之和等于1。变换前后的空间关系见图1-3-1。

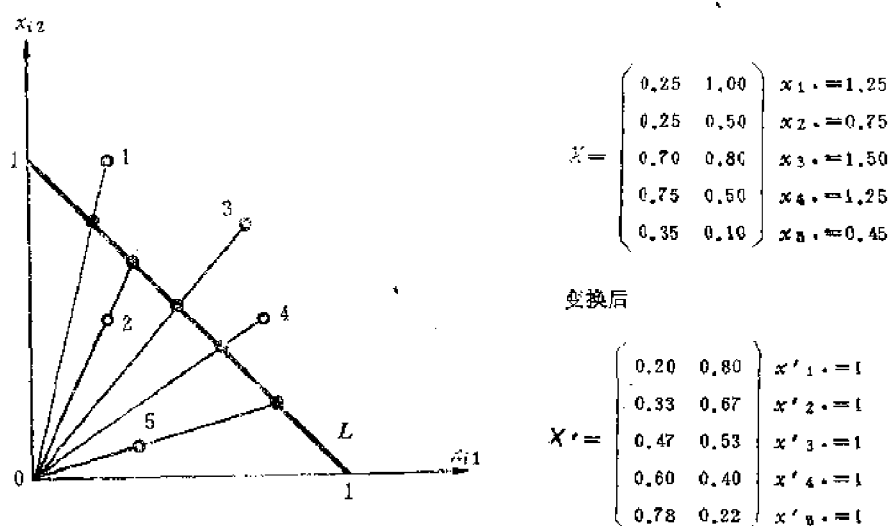


图1-3-1　二维的总和标准化变换

二、最大值标准化

对变量进行最大值标准化变换时，是将变量的各样品观测值除以该变量中的最大样品观测值。即：

$$x'_{ij}=\frac{x_{ij}}{\max\limits_{1\leqslant j\leqslant n} x_{ij}}\qquad (i=1,2,\cdots,m;\ j=1,2,\cdots,n) \qquad (1\text{-}3\text{-}3)$$

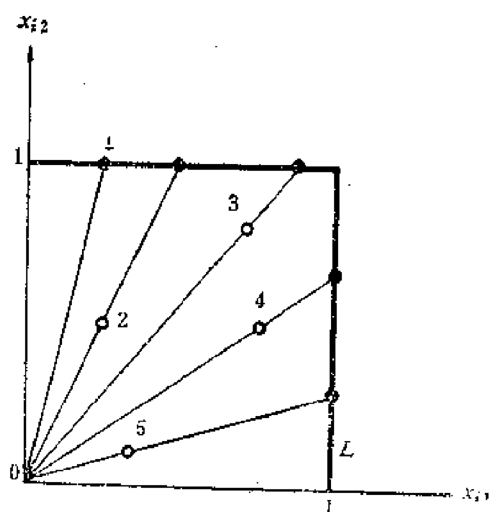
对表1-3-1中的变量进行最大值标准化变换时，首先要找出矩阵中每行的最大值，变换后得到：

$$X=\begin{pmatrix} 1000 & 250 & 100 & 10 & 40 \\ 500 & 150 & 70 & 200 & 100 \\ 1.5 & 1 & 3 & 2 & 5 \\ 2000 & 2200 & 1500 & 1800 & 2500 \end{pmatrix}\begin{matrix} \max(x_{1\cdot})=1000 \\ \max(x_{2\cdot})=500 \\ \max(x_{3\cdot})=5 \\ \max(x_{4\cdot})=2500 \end{matrix}$$

$$X'=\begin{pmatrix} 1 & 0.25 & 0.10 & 0.01 & 0.04 \\ 1 & 0.30 & 0.14 & 0.40 & 0.20 \\ 0.30 & 0.20 & 0.60 & 0.40 & 1 \\ 0.80 & 0.88 & 0.60 & 0.72 & 1 \end{pmatrix}\begin{matrix} \max(x'_{1\cdot})=1 \\ \max(x'_{2\cdot})=1 \\ \max(x'_{3\cdot})=1 \\ \max(x'_{4\cdot})=1 \end{matrix}$$

经变换，每个变量的最大样品值为1，其余的样品值都是小数。

在二维情况下，对变量进行总和标准化变换时，在空间关系上是将原变量点径向投影到单位方阵上。见图1-3-2。



$$X=\begin{pmatrix} 0.25 & 1.00 \\ 0.25 & 0.50 \\ 0.70 & 0.80 \\ 0.75 & 0.50 \\ 0.35 & 0.10 \end{pmatrix} \begin{matrix} \max(x_{1\cdot})=1.00 \\ \max(x_{2\cdot})=0.50 \\ \max(x_{3\cdot})=0.80 \\ \max(x_{4\cdot})=0.75 \\ \max(x_{5\cdot})=0.35 \end{matrix}$$

变换后

$$X'=\begin{pmatrix} 0.25 & 1 \\ 0.50 & 1 \\ 0.875 & 1 \\ 1 & 0.67 \\ 1 & 0.26 \end{pmatrix} \begin{matrix} \max(x'_{1\cdot})=1 \\ \max(x'_{2\cdot})=1 \\ \max(x'_{3\cdot})=1 \\ \max(x'_{4\cdot})=1 \\ \max(x'_{5\cdot})=1 \end{matrix}$$

图1-3-2　二维的最大值标准化变换

三、模 标 准 化

对变量进行模标准化变换时，是把每个变量看作是一个n维样品空间的向量$\vec{X}_i(i=1,2,\cdots,m)$，向量的长度，即模为：

$$|\vec{X}_i|=\sqrt{\sum_{j=1}^{n}x_{ij}^2} \qquad (i=1,2,\cdots,m) \tag{1-3-4}$$

变换时，是将变量的各样品观测值除以该变量的模，即：

$$x'_{ij}=\frac{x_{ij}}{|\vec{X}_i|} \qquad (i=1,2,\cdots,m;\ j=1,2,\cdots,n) \tag{1-3-5}$$

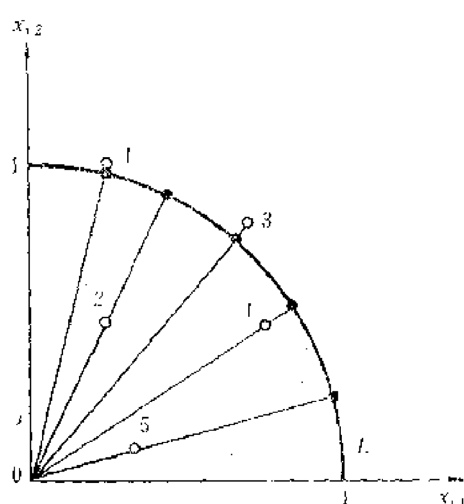
对表1-3-1中的变量进行模标准化变换时，首先要算出矩阵中每行的模，变换后得到：

$$X=\begin{pmatrix} 1000 & 250 & 100 & 10 & 40 \\ 500 & 150 & 70 & 200 & 100 \\ 1.5 & 1 & 3 & 2 & 5 \\ 2000 & 2200 & 1500 & 1800 & 2500 \end{pmatrix} \begin{matrix} |\vec{X}_1|=1036.436 \\ |\vec{X}_2|=572.189 \\ |\vec{X}_3|=6.423 \\ |\vec{X}_4|=4536.518 \end{matrix}$$

$$X'=\begin{pmatrix} 0.965 & 0.241 & 0.096 & 0.010 & 0.039 \\ 0.847 & 0.262 & 0.122 & 0.350 & 0.175 \\ 0.234 & 0.156 & 0.467 & 0.311 & 0.778 \\ 0.441 & 0.485 & 0.331 & 0.397 & 0.551 \end{pmatrix} \quad \sum_{j=1}^{5}(x'_{ij})^2=1$$

变换后的特征是每个变量的平方和等于1。

在二维情况下，对变量进行模标准化变换时，在空间关系上是将原变量点径向投影到单位弧L上。见图1-3-3。



$$X=\begin{pmatrix} 0.25 & 1.00 \\ 0.25 & 0.50 \\ 0.70 & 0.80 \\ 0.75 & 0.50 \\ 0.35 & 0.10 \end{pmatrix} \quad \begin{array}{l} |\vec{X}_1|=1.031 \\ |\vec{X}_2|=0.559 \\ |\vec{X}_3|=1.063 \\ |\vec{X}_4|=0.901 \\ |\vec{X}_5|=0.364 \end{array}$$

变换后

$$X'=\begin{pmatrix} 0.243 & 0.970 \\ 0.447 & 0.894 \\ 0.659 & 0.753 \\ 0.832 & 0.555 \\ 0.962 & 0.275 \end{pmatrix} \quad \sum_{i=1}^{2}(x'_{ij})^2=1$$

图 1-3-3 二维的模标准化变换

四、中心标准化

对变量进行中心标准化变换时，是把变量的各样品观测值减去该变量的平均值。即：

$$x'_{ij}=x_{ij}-\bar{x}_i \qquad (1\text{-}3\text{-}6)$$

$$(i=1,\ 2,\ \cdots m;\ j=1,\ 2,\ n)$$

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij} \qquad (1\text{-}3\text{-}7)$$

$$(i=1,\ 2,\ \cdots,\ m)$$

对表1-3-1中的变量进行中心标准化变换时，首先要计算矩阵中每行的平均值，变换后得到

$$X=\begin{pmatrix} 1000 & 250 & 100 & 10 & 40 \\ 500 & 150 & 70 & 200 & 100 \\ 1.5 & 1 & 3 & 2 & 5 \\ 2000 & 2200 & 1500 & 1800 & 2500 \end{pmatrix} \begin{array}{l} \bar{x}_1=280 \\ \bar{x}_2=204 \\ \bar{x}_3=2.5 \\ \bar{x}_4=2000 \end{array}$$

$$X'=\begin{pmatrix} 720 & -30 & -180 & -270 & -240 \\ 296 & -54 & -134 & -4 & -104 \\ -1 & -1.5 & 0.5 & -0.5 & 2.5 \\ 0 & 200 & -500 & -200 & 500 \end{pmatrix} \quad \sum_{j=1}^{5}x'_{ij}=0$$

变换后的特征是每个变量的总和等于0。

在二维情况下，对变量进行中心标准化变换时，在空间关系上是将原变量点按法线方向投影到斜率为－1的直线上。见图1-3-4。

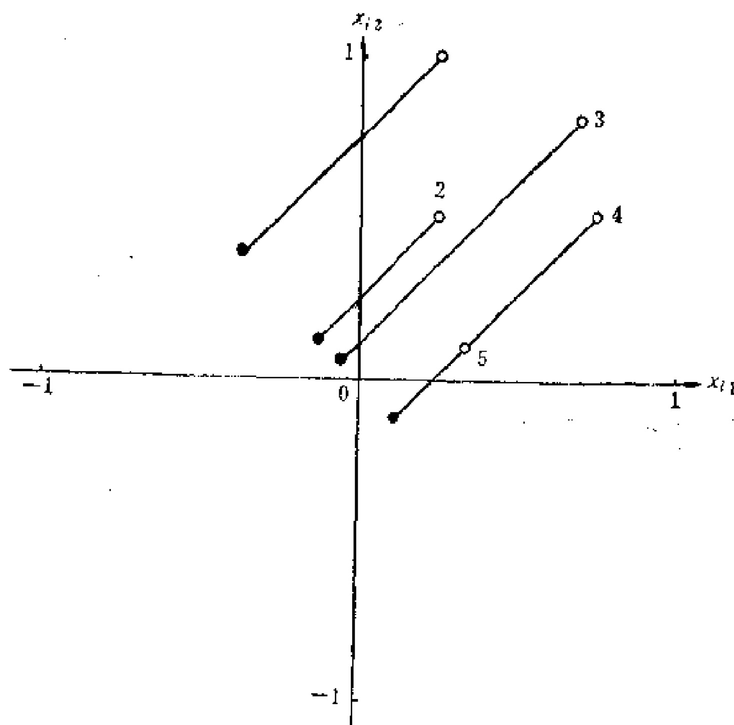


图1-3-4　二维的中心标准化变换

$$X=\begin{pmatrix} 0.25 & 1.00 \\ 0.25 & 0.50 \\ 0.70 & 0.80 \\ 0.75 & 0.50 \\ 0.35 & 0.10 \end{pmatrix}\begin{matrix} \bar{x}_1=0.625 \\ \bar{x}_2=0.375 \\ \bar{x}_3=0.750 \\ \bar{x}_4=0.625 \\ \bar{x}_5=0.225 \end{matrix} \quad \text{变换后} \quad X'=\begin{pmatrix} -0.375 & 0.375 \\ -0.125 & 0.125 \\ -0.050 & 0.050 \\ 0.125 & -0.125 \\ 0.125 & -0.125 \end{pmatrix} \quad \sum_{i=1}^{2} x'_{ij}=0$$

五、标准差标准化

对变量进行标准差标准化变换时，是将变量的各样品观测值减去该变量的平均值，所得之差除以该变量的标准差。即：

$$x'_{ij}=\frac{x_{ij}-\bar{x}_i}{s_i} \qquad (i=1,2\cdots,m;\ j=1,2,\cdots,n) \tag{1-3-8}$$

其中

$$\bar{x}_i=\frac{1}{n}\sum_{j=1}^{n} x_{ij}, \qquad (i=1,2,\cdots,m) \tag{1-3-9}$$

$$s_i=\sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2} \qquad (i=1,2,\cdots,m) \tag{1-3-10}$$

对表1-3-1中的变量进行标准差标准化变换时，首先要计算矩阵中每行的平均值及标准差，变换后得到

$$X=\begin{pmatrix} 1000 & 250 & 100 & 10 & 40 \\ 500 & 150 & 70 & 200 & 100 \\ 1.5 & 1 & 3 & 2 & 5 \\ 2000 & 2200 & 1500 & 1800 & 2500 \end{pmatrix} \begin{matrix} \bar{x}_1=280 & s_1=369.378 \\ \bar{x}_2=204 & s_2=154.480 \\ \bar{x}_3=2.5 & s_3=1.414 \\ \bar{x}_4=2000 & s_4=340.588 \end{matrix}$$

$$X'=\begin{pmatrix} 1.949 & -0.081 & -0.487 & -0.731 & -0.650 \\ 1.916 & -0.350 & -0.867 & -0.026 & -0.673 \\ -0.707 & -1.061 & 0.354 & -0.354 & 1.768 \\ 0 & 0.587 & -1.468 & -0.587 & 1.468 \end{pmatrix} \quad \bar{x}'_i=0 \qquad s'_i=1$$

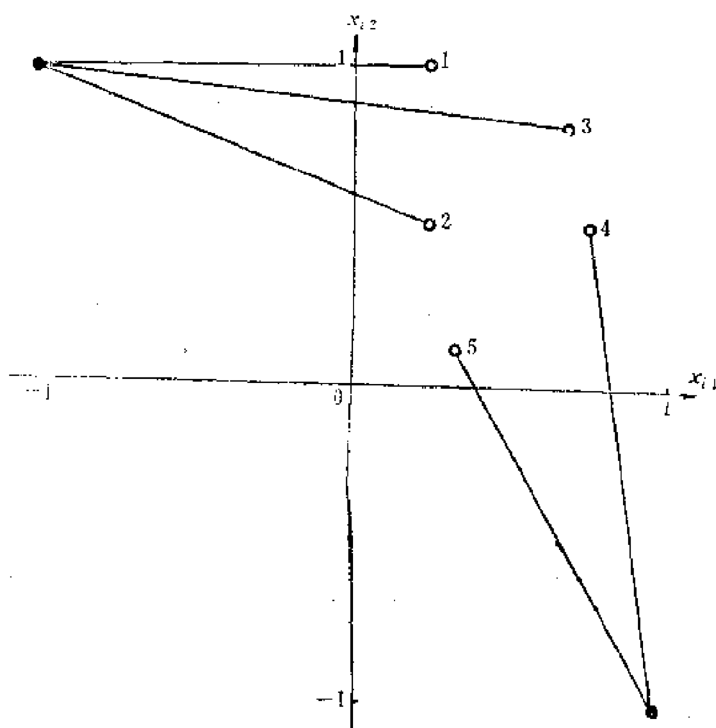


图1-3-5　二维的标准差标准化变换

$$X=\begin{pmatrix} 0.25 & 1.00 \\ 0.25 & 0.50 \\ 0.70 & 0.80 \\ 0.75 & 0.50 \\ 0.35 & 0.10 \end{pmatrix} \begin{matrix} \bar{x}_1=0.625 & s_1=0.375 \\ \bar{x}_2=0.375 & s_2=0.125 \\ \bar{x}_3=0.750 & s_3=0.050 \\ \bar{x}_4=0.625 & s_4=0.125 \\ \bar{x}_5=0.225 & s_5=0.125 \end{matrix} \quad \text{变换后} \quad X'=\begin{pmatrix} -1 & 1 \\ -1 & 1 \\ -1 & 1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \bar{x}'_i=0 \quad s'_i=1$$

标准差标准化变换后的特征是所有变量的平均值都等于0，标准差都等于1。因此，经过标准差标准化变换后的数据，被称为规格化数据。

在二维情况下，对变量进行标准差标准化变换时，在空间关系上是将原变量点分别投影到（-1，1）或（1，-1）两个点上去。见图1-3-5。

六、极差标准化

对变量进行极差标准化变换时，是将变量的各样品观测值减去该变量的平均值，所得之差除以该变量的极差。即：

$$x'_{ij}=\frac{x_{ij}-\bar{x}_i}{\max\limits_{1\leq j\leq n}x_{ij}-\min\limits_{1\leq j\leq n}x_{ij}} \qquad (i=1,2,\cdots,m;\ j=1,2,\cdots,n) \qquad (1\text{-}3\text{-}11)$$

对表1-3-1中的变量进行极差标准化变换时，首先要计算矩阵中每行的平均值及极差，这里令极差 $\max\limits_{1\leq j\leq n}x_{ij}-\min\limits_{1\leq j\leq n}x_{ij}=Lx_i$，变换后得到

$$X=\begin{pmatrix} 1000 & 250 & 100 & 10 & 40 \\ 500 & 150 & 70 & 200 & 100 \\ 1.5 & 1 & 3 & 2 & 5 \\ 2000 & 2200 & 1500 & 1800 & 2500 \end{pmatrix} \begin{matrix} \bar{x}_1=280 & Lx_1=990 \\ \bar{x}_2=204 & Lx_2=430 \\ \bar{x}_3=2.5 & Lx_3=4 \\ \bar{x}_4=2000 & Lx_4=1000 \end{matrix}$$

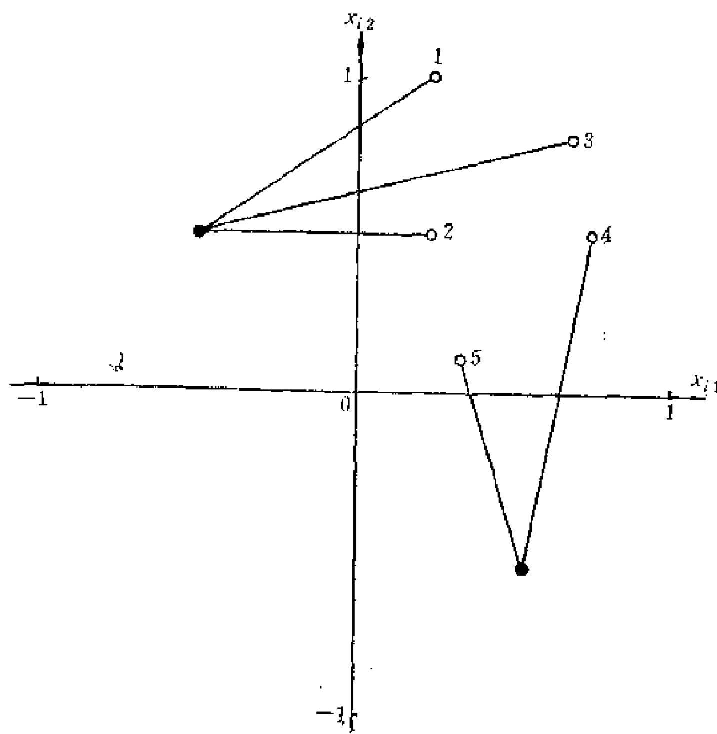


图1-3-6 二维的极差标准化变换

$$X=\begin{pmatrix} 0.25 & 1.00 \\ 0.25 & 0.50 \\ 0.70 & 0.80 \\ 0.75 & 0.50 \\ 0.35 & 0.10 \end{pmatrix} \begin{matrix} \bar{x}_1=0.625 & Lx_1=0.75 \\ \bar{x}_2=0.375 & Lx_2=0.25 \\ \bar{x}_3=0.750 & Lx_3=0.10 \\ \bar{x}_4=0.625 & Lx_4=0.25 \\ \bar{x}_5=0.225 & Lx_5=0.25 \end{matrix} \quad 变换后 \quad X'=\begin{pmatrix} -0.5 & 0.5 \\ -0.5 & 0.5 \\ -0.5 & 0.5 \\ 0.5 & -0.5 \\ 0.5 & -0.5 \end{pmatrix} \quad Lx_i=1$$

$$X'=\begin{pmatrix} 0.727 & -0.030 & 0.182 & -0.273 & -0.242 \\ 0.688 & -0.126 & -0.312 & -0.009 & -0.242 \\ -0.250 & -0.375 & 0.125 & -0.125 & 0.625 \\ 0 & 0.200 & -0.500 & -0.200 & 0.500 \end{pmatrix} \quad Lx_i=1$$

极差标准化变换后的特征是每个变量的极差都等于1。

在二维情况下，对变量进行极差标准化变换时，在空间关系上是将原变量点分别投影到（0.5，−0.5）或（−0.5，0.5）两个点上去。见图1-3-6。

七、极差正规化

对变量进行极差正规化变换时，是将每个变量的各样品观测值减去该变量的极小值，所得之差除以该变量的极差。即：

$$x'_{ij}=\frac{x_{ij}-\min\limits_{1\leqslant j\leqslant n} x_{ij}}{\max\limits_{1\leqslant j\leqslant n} x_{ij}-\min\limits_{1\leqslant j\leqslant n} x_{ij}} \qquad (i=1,2,\cdots,m;\ j=1,2,\cdots,n) \qquad (1\text{-}3\text{-}12)$$

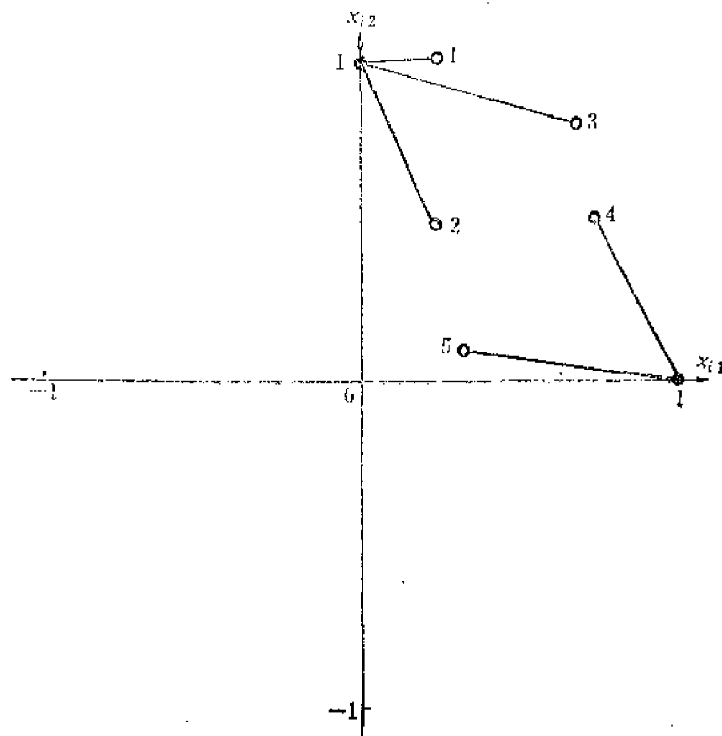


图1-3-7　二维的极差正规化变换

$$X=\begin{pmatrix} 0.25 & 1.00 \\ 0.25 & 0.50 \\ 0.70 & 0.80 \\ 0.75 & 0.50 \\ 0.35 & 0.10 \end{pmatrix} \begin{matrix} \min x_1=0.25 & Lx_1=0.75 \\ \min x_2=0.25 & Lx_2=0.25 \\ \min x_3=0.70 & Lx_3=0.10 \\ \min x_4=0.50 & Lx_4=0.25 \\ \min x_5=0.10 & Lx_5=0.25 \end{matrix} \quad 变换后\ X'=\begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \quad 0\leqslant x'_{ij}\leqslant 1$$

对表1-3-1中的变量进行极差正规化变换时，首先要计算矩阵中每行的最小值及极差，这里令极小值 $\min\limits_{1<j<n} x_{ij}=\min x_i$，变换后得到

$$X=\begin{pmatrix} 1000 & 250 & 100 & 10 & 40 \\ 500 & 150 & 70 & 200 & 100 \\ 1.5 & 1 & 3 & 2 & 5 \\ 2000 & 2200 & 1500 & 1800 & 2200 \end{pmatrix} \begin{matrix} \min x_1=10 & Lx_1=990 \\ \min x_2=70 & Lx_2=430 \\ \min x_3=1 & Lx_3=4 \\ \min x_4=1500 & Lx_4=1000 \end{matrix}$$

$$X'=\begin{pmatrix} 1 & 0.242 & 0.091 & 0 & 0.030 \\ 1 & 0.186 & 0 & 0.302 & 0.070 \\ 0.125 & 0 & 0.500 & 0.250 & 1 \\ 0.500 & 0.700 & 0 & 0.300 & 1 \end{pmatrix} \qquad 0\leqslant x'_{ij}\leqslant 1$$

极差正规化变换后的特征是每个变量的值均在0与1之间，其中一个最大值为1，一个最小值为0，即变换到 [0，1] 闭区间范围内。

在二维情况下，对变量进行极差正规化变换时，是将原变量点分别投影到（1，0）或（0，1）两个点上去。见图1-3-7。

第二节 定性数据的定量化变换

地质数据中有一些数据属于定性数据，在地质研究工作中也要经常用到定性数据，有时甚至是不可避免的。例如，生油岩的颜色，钻井取出的岩心或岩屑的含油状态，地层中某种古生物化石的有无等等都是不能用数值描述的定性数据。定性数据不能直接参加运算，为了使定性数据能够用于定量研究，必须将定性数据的符号或代码赋以定量的数值，这就是对定性数据的定量化变换。近年来，对定性数据的预处理，虽然已引起数学地质界的重视，然而，研究的深度还很不够。

一、二态定性数据的变换

如果名义型数据或有序型数据只有两种对立状态，则称为二态定性数据。两种对立状态是指在两种仅有的状态中必须是其中的一种，亦即所谓"非此即彼"。在这种情况下，对定性数据采用0，1化变换是十分方便的。

二态定性数据的0，1化变换，其具体作法是把两种对立状态中的一种状态赋值为1，另一种状态赋值为0。通常是将对所研究问题有利的或肯定的状态赋值为1，将不利的或否定的状态赋值为0。

例如，进行某一地区的地层对比时，同一层位的不同观测点，有时可找到对分层有意义的某种古生物化石，有时找不到这种古生物化石。如果进行数据变换时，对出现的地点可赋值为1，不出现的地点可赋值为0。又如，在钻井过程中进行岩屑录井时，对于出现油砂的层位可赋值为1，不出现油砂的层位可赋值为0。在一般的情况下，

状态	有利或肯定状态	不利或否定状态
赋值	1	0

在此需要指出，0，1化是一种简单而实用的变换方法。而且变换后的0，1化数据可与极差正规化变换后的定量数据混合使用，前已述及，经极差正规化变换后的定量数据，其特征是所有数据均被压缩到［0，1］闭区间范围内，因而，经0，1化变换后的定性数据可看作是定量数据的两种极端状态。可见，这种变换方法是符合实际情况的。

二、有序型多态定性数据的变换

如果定性数据的状态数是有限的，而且状态可按一定的次序进行排列，这种数据可称为有序型多态定性数据。

例如，从钻井中所取出的岩心，按其含油程度可分为如下四个级别，即：

状态	不含油	油斑	含油	饱含油
赋值	0	1	2	3

对有序型多态定性数据进行等级变换时，一般是用非负整数对状态进行赋值，由最低级的状态到最高级的状态，其赋值要逐渐增大。根据实际情况，可采用等差式的等级赋值，也可以采用非等差式的等级赋值。上面谈到对岩心含油程度的赋值就是等差式的等级赋值，极差为1。

又如，根据泥岩的颜色划分生油条件时，采用下面的赋值方法：

状态	红色	浅灰色	灰色	黑色
赋值	0	1	3	6

这种赋值就是非等差式的赋值方法。

有序多态定性数据经过这种变换后，如果再按前一节中定量数据的变换方法作进一步的变换，则变换后的多态定性数据可与定量数据混合使用。

第三节　非线性数据的线性变换

一种地质数据与另一种地质数据之间的关系，多数情况下是非线性关系。为了计算、处理或图形表现上的方便，经常需要把这种非线性关系转化为线性关系。对于这种变换有时也称作曲线化直线的变换。

在地质研究工作中，两种地质变量间经常碰到的非线性关系，有如下一些函数类型。

一、幂　函　数

幂函数的表达式为

$y=ax^b \quad (a>0)$

其中x为自变量，y为因变量，幂函数的曲线图形见图1-3-8。

变换时，可令 $X=\lg x$，$Y=\lg y$，则有

$$Y=\lg a+bX \tag{1-3-13}$$

此处要求　$x>0$，$y>0$，$a>0$。

变换后，在双对数坐标纸上（1-3-13）式为一条直线。

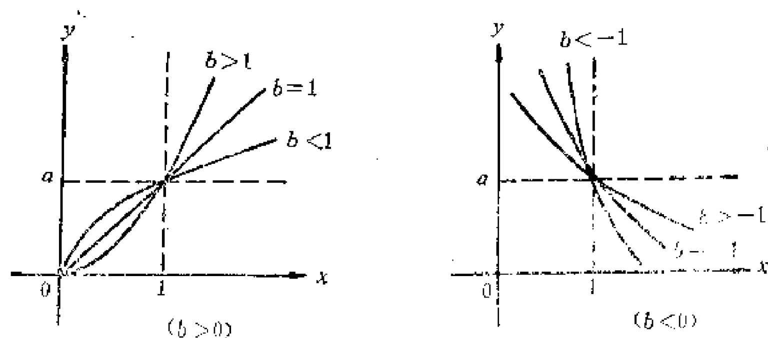


图1-3-8

二、指数函数

指数函数的表达式为

$$y=ae^{bx}\quad (a>0)$$

指数函数的曲线图形见图1-3-9。

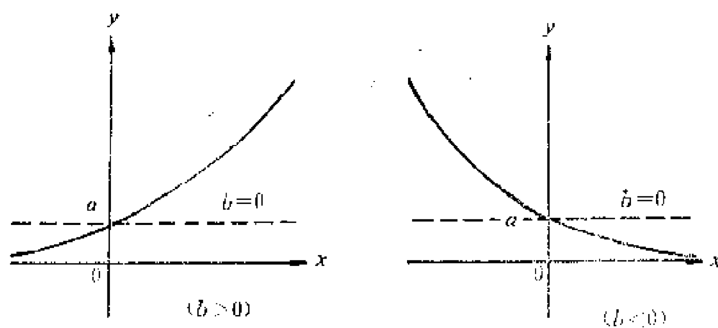


图1-3-9

变换时，可令 $X=x$，$Y=\lg y$，则有

$$Y=\lg a+(b\lg e)X \tag{1-3-14}$$

此处要求 $y>0$，$a>0$。

变换后，在单对数（Y 轴）坐标纸上（1-3-14）式为一条直线。

三、对数函数

对数函数的表达式为

$$y=a+b\lg x$$

对数函数的曲线图形见图1-3-10。

变换时，可令$X=\lg x$，$Y=y$，则有

$$Y=a+bX \tag{1-3-15}$$

此处要求　$x>0$。

变换后，在单对数（X轴）坐标纸上（1-3-15）式为一条直线。

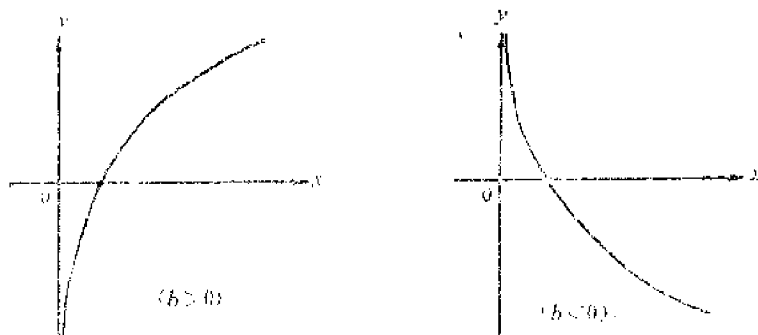


图1-3-10

四、 $y=ae^{\frac{b}{x}}$

函数的曲线图形见图1-3-11。

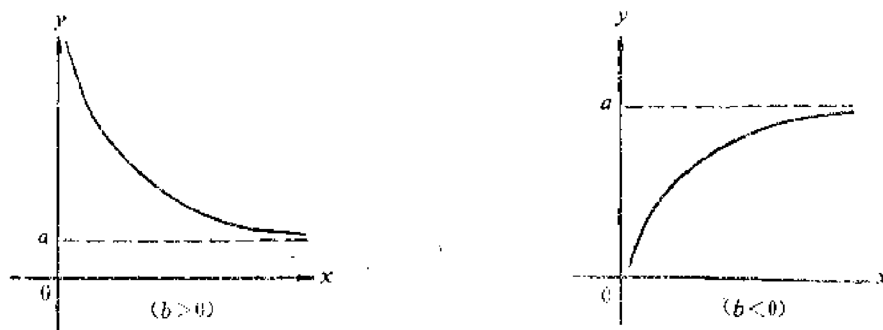


图1-3-11

变换时，可令$X=\frac{1}{x}$，$Y=\lg y$，则有

$$Y=\lg a+(b\lg e)X \tag{1-3-16}$$

此处要求　$y>0$，$a>0$，$x\neq 0$。

变换后，在单对数（Y轴）坐标纸上（1-3-16）式为一条直线。

五、 $y=\frac{1}{a+be^{-x}}$　（$a>0$）

函数的曲线图形见图1-3-12。

变换时，可令$X=e^{-x}$，$Y=\frac{1}{y}$，则有

$$Y=a+bX \tag{1-3-17}$$

此处要求　$y\neq 0$。

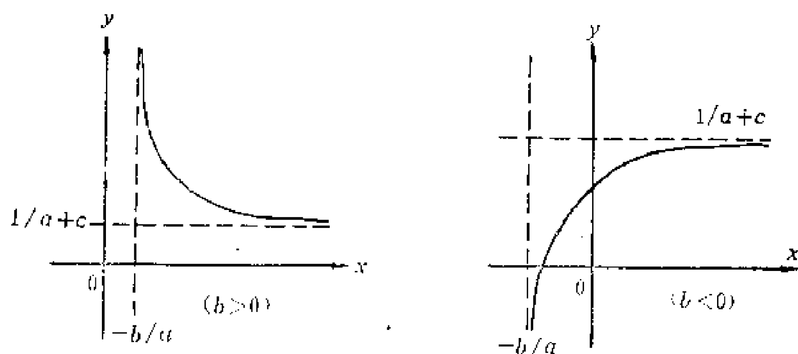


图1-3-12

六、$y=\frac{x}{ax+b}+c$　（$a>0$；$c>0$）

函数的曲线图形见图1-3-13。

变换时，可令$X=\frac{1}{x}$，$Y=\frac{1}{y-c}$，则有

$$Y=a+bX \qquad (1\text{-}3\text{-}18)$$

此处要求　$x\neq 0$，$y-c\neq 0$。

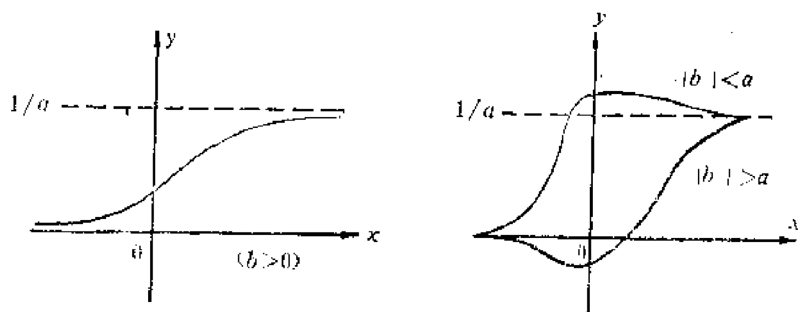


图1-3-13

七、$y=\frac{1}{ax+b}$（$a>0$）

函数的曲线图形见图1-3-14。

变换时，可令$X=x$，$Y=\frac{1}{y}$，则有

$$Y=b+aX \qquad (1\text{-}3\text{-}19)$$

此处要求　$y\neq 0$。

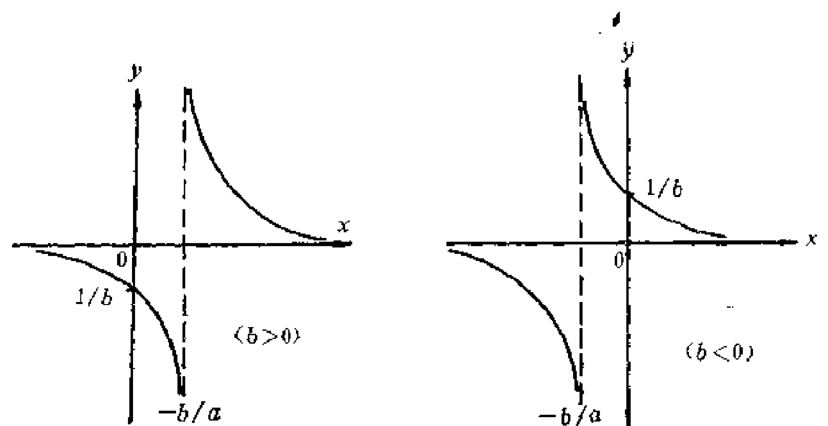


图1-3-14

八、$y=\dfrac{ax+b}{cx+d}$　$\left(D=\begin{vmatrix} a & b \\ c & d \end{vmatrix}\neq 0\right)$

函数的曲线图形见图1-3-15。

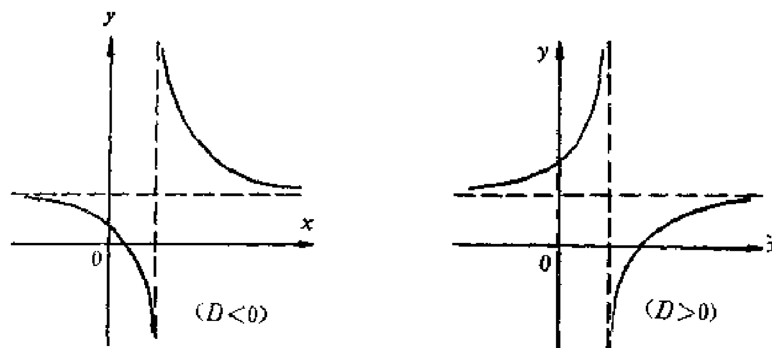


图1-3-15

变换时，首先要在曲线上选取一点（x_0，y_0），并且令：

$$X=x,\ Y=\frac{x-x_0}{y-y_0}$$

则有　$Y=A+BX$　　（1-3-20）

最后用回归方法，可由已给出的数据确定A和B。

第四节　原始数据的简缩与增补

当地质数据的样品数量很多很多时，将会使计算量大大增加，甚至在计算过程中出现病状问题。因此，需要对那些作用不大、可有可无的多余数据予以舍弃，这就是数据的简缩。

在绝大多数情况下，探区中各处投入的勘探工作量是不均匀的，特别是早期勘探阶段尤其如此。此外，探区中各处所配置的勘探工种也不会完全一致，所以所获得的数据内容也就不完全一样。如果将所有数据项汇总在一张综合表上，则会出现地质数据项目不全的问题，

在这些情况下则需要补充数据，这就是数据的增补。

一、数据简缩的方法

1. 分区加权法

如果某项地质数据的数量非常大时，或者是数据在区域上的分布极不均匀时，为了避免计算时间过长，可以采用分区加权法对原始数据进行预处理。

假如在一个探区中有N个地质数据，则可根据实际需要将探区分成大小相等或不相等的m个小区，要求在每个小区中至少有一个数据点。如果其中第j个小区中有n_j个数据点，那么则有：

$$N=n_1+n_2+\cdots+n_j+\cdots+n_m$$

这里令第j个小区中每个数据点的权为$\frac{1}{n_j}$，则每个小区中所包含的数据点的权和等于1。因而有

$$n_1\frac{1}{n_1}+n_2\frac{1}{n_2}+\cdots+n_j\frac{1}{n_j}+\cdots+n_m\frac{1}{n_m}=m$$

这样一来，则每个小区相当于一个有效的数据点，这就将原来数据量很大的地质数据简化为m个有效数据点。在尔后的计算只要用m个有效数据点就可以了。

地质数据经常是多变量数据，如果每个数据由p个地质变量组成，则第j个小区的第k个地质变量的简缩值可用（1-3-21）式表示，即

$$z_{kj}=\frac{1}{n_j}\sum_{i=1}^{n_j}z_{kji} \quad (j=1,2,\cdots,m;\ k=1,2,\cdots,p) \tag{1-3-21}$$

式中 z_{kj}——第k个地质变量的第j个小区的简缩值；

n_j——第j个小区中的地质数据个数；

z_{kji}——第j个小区中第i个数据的第k个地质变量。

2. 分区滑动平均法

分区滑动平均法与分区加权法一样，也要将研究区分成若干个小区，分区原则二者相同。但是，分区滑动平均法要考虑简缩后的数据点位置。

如果第j个小区中有n_j个数据点，每个数据点有p个地质变量，其中第i个地质数据的坐标为（x_{kji}，y_{kji}），变量数值为z_{kji}。第j个小区简缩后的有效数据点的坐标值及变量值可用下面的（1-3-22）、（1-3-23）、（1-3-24）式计算求出。

$$x_{kj}=\sum_{i=1}^{n_j}x_{kji}\cdot z_{kji}\Big/\sum_{i=1}^{n_j}z_{kji} \tag{1-3-22}$$

$$y_{kj}=\sum_{i=1}^{n_j}y_{kji}\cdot z_{kji}\Big/\sum_{i=1}^{n_j}z_{kji} \tag{1-3-23}$$

$$z_{kj}=\sum_{i=1}^{k_j}z_{kji}\Big/n_j \quad (j=1,2,\cdots,m;\ k=1,2,\cdots,p) \tag{1-3-24}$$

（1-3-22）、（1-3-23）、（1-3-24）式中

x_{kj}、y_{kj}——分别为第k个地质变量的第j个小区简缩后的横坐标与纵坐标；

z_{kj}——第k个地质变量的第j个小区的简缩值；

x_{kji}、y_{kji}——分别为第k个地质变量的第j个小区中第i个数据的横坐标与纵坐标；

z_{kji}——第k个地质变量的第j个小区的第i个观测值；

n_j——第j个小区中的地质数据个数。

3. 随机删点法

如果探区中某些局部地区的数据点过密，则可以随机地删去一些数据点，这不仅可以减少计算工作量，也可以保证计算过程的稳定性。

随机删点时，可以人为地随机删除一些数据点；也可以对数据点顺序编号，按随机数抽样法进行删除。随机数抽样法详见本章第八节。

二、数据的增补方法

在没有数据点的大片空白地区，为了补充一些数据点，可以用临近已有的地质数据按外推法，即根据数据的变化趋势，补充一些数量适当的数据点；或者根据临区的已有数据点，用某种约定的插值方法补充一些数量适当的数据点。

但是，必须说明，补点的目的是为了全区计算上的稳定，而补点后原空白区的计算结果是不可信的。

此外，由于探区中各处取样点的化验分析项目不完全一致，因而使多变量的数据矩阵中，某些样品缺少一些变量的数值。但是，由于研究工作上的需要又不能删掉这些缺少数值的变量。为了增补缺项位置上的变量数值，一般情况下可用该变量已有数据的平均值代替或者用区域上临近数据的平均值代替。

第五节　混合数据的预处理

混合数据是指定量数据和定性数据兼而有之的数据集合。当同时使用定量数据和定性数据研究某个地质问题时，可以采用下面简单易行的方法进行预处理。这种方法的要点是建立混合数据之间的相似系数矩阵，而建立矩阵前首先需要把每个变量的数值变换到0—1之间的范围内。

例如，有n个样品，每个样品有m个变量，m个变量中有m_1个定量数据，有m_2个定性数据，即：$m_1+m_2=m$。

这里令样品X_i与样品X_j之间的相似系数为r_{ij}，即

$$r_{ij}=\frac{1}{m}\sum_{k=1}^{m}r_{ijk} \quad (i,j=1,2,\cdots,n;\ k=1,2,\cdots,m) \tag{1-3-25}$$

式中　r_{ijk}——为样品X_i与X_j之间的第k个变量的相似性指标。

如果第k个变量为定量数据时，可规定：

$$r_{ijk}=x_{ik}\cdot x_{jk} \tag{1-3-26}$$

式中　x_{ik}、x_{jk}——分别是样品X_i与X_j的第k个变量的数值，这里要求将其数值变换到[0,1]或者（0，1）区间范围内。

如果第k个变量为定性数据时，可规定：

$$r_{ijk}=\begin{cases}0 & （X_i与X_j的状态不同时） \\ x_k\quad (0\leqslant x_k\leqslant 1) & （X_i与X_j的状态相同时）\end{cases}\qquad (1\text{-}3\text{-}27)$$

式中 x_k——当X_i与X_j二者状态相同时的取值，取值大小可视变量的重要性而定。

为了便于说清混合数据的预处理过程，这里给出一个实际算例。

某个地质凹陷有3个地质圈闭，每个圈闭有5项地质变量，其中前3项为定量数据，后2项为定性数据，见表1-3-2。

表1-3-2 地质圈闭数据表

变量＼样品	X_1	X_2	X_3
闭合面积(10^2m^2)	200	150	70
闭合度(m)	100	200	40
埋藏深度(m)	2000	1700	1500
有无断层	无	有	有
有无火成岩侵入	无	无	有

表1-3-2中的定量数据可用极差正规化变换为［0，1］之间的小数；也可以用总和标准化变换为（0，1）之间的小数。这里是采用总和标准化方法处理的。定性数据可按二态定性数据0，1化变换为0或1。变换后，有如下矩阵：

$$X=\begin{pmatrix}0.467 & 0.375 & 0.167\\ 0.294 & 0.588 & 0.118\\ 0.385 & 0.327 & 0.288\\ 1 & 0 & 0\\ 1 & 1 & 0\end{pmatrix}$$

考虑到无断层、无火成岩侵入这种状态对于油气形成有利，所以赋值为1；而有断层、有火成岩侵入状态对油气形成不利，因而赋值为0。

这里规定定性数据样品间的状态相同时，$r_{ijk}=0.5$，状态不同时$r_{ijk}=0$。定量数据间的相似系数是按（1-3-25）式计算。例如：第一个样品与本身的相似系数r_{11}为

$$r_{11}=\frac{1}{5}(0.467^2+0.294^2+0.385^2+0.5+0.5)=0.292$$

又如第二个样品与第三个样品之间的相似系数为

$$r_{23}=\frac{1}{5}(0.375\times0.167+0.588\times0.118+0.327\times0.288+0.5+0)=0.145$$

同样，其他各样品间的相似系数也可依此计算，最后可以得到所有样品之间的相似系数，即有如下相似系数矩阵：

$$R=[r_{ij}]_{3\times 3}=\begin{pmatrix} 0.292 & 0.194 & 0.045 \\ 0.194 & 0.316 & 0.145 \\ 0.045 & 0.145 & 0.225 \end{pmatrix}$$

此时，虽然样品间的相似系数在（0，1）之间，但是，在此矩阵中每个样品与其自身的相似系数并不等于1；而且也不同于距离系数，因为样品与其自身的距离系数应等于0。

为了构成样品间的距离系数矩阵D，使其样品与自身的距离$d_{ii}=0$，可对矩阵中的每个元素作如下变换：

$$d_{ij}=r_{ii}+r_{jj}-2r_{ij} \qquad (i,j=1,2,\cdots,n) \qquad (1\text{-}3\text{-}28)$$

则可得到样品间的距离系数矩阵D

$$D=[d_{ij}]_{3\times 3}=\begin{pmatrix} 0 & 0.220 & 0.427 \\ 0.220 & 0 & 0.251 \\ 0.427 & 0.251 & 0 \end{pmatrix}$$

为了构成样品间的相似系数矩阵Q，使其样品与自身的相似性$q_{ii}=1$，可对矩阵作进一步的变换：

$$q_{ij}=1-d_{ij}=1-(r_{ii}+r_{jj}-2r_{ij}) \qquad (1\text{-}3\text{-}29)$$

$$(i,\ j=1,\ 2,\ \cdots,\ n)$$

则可以得到样品间的相似系数矩阵Q

$$Q=[q_{ij}]_{3\times 3}=\begin{pmatrix} 1 & 0.780 & 0.573 \\ 0.780 & 1 & 0.749 \\ 0.573 & 0.749 & 1 \end{pmatrix}$$

经过上述预处理之后，便可以进行样品分类或其他方面的多元统计分析计算。

第六节　离群数据的处理

地质数据中经常出现为数极少的特高值数据或特低值数据，这种数据有时比数据集合的平均值高出很多倍或低很多倍。这些为数极少的数据，通常称作数据集合中的离群数据，或奇异数据。有时把其中的特高值数据专称为风暴数据。

目前，对离群数据的处理有两种途径。其一是修改数据，适应方法。也就是说用某种方法将离群数据经过预处理之后再参加计算；或者用某种方法首先找出离群数据，将其舍弃而根本不参加计算。其二是修改方法，适应数据。也就是说通过改善目前的计算方法，虽然保留离群数据，但使离群数据不起干扰作用或少起干扰作用。例如，近年发展起来的稳健统计学方法就属于后一种途径。而本节讲述的内容都属于前一种途径。

离群数据中，虽然包括特高值数据，也包括特低值数据，但是，多数情况下离群数据是指特高值数据而言。在地质研究工作中，对特低值的舍弃一般无争议；而对特高值的舍弃，地质家们一般是很慎重的，甚至不愿意轻易舍弃。因而，首要的问题是要确定一个离群数据

的界线，才能合理地进行处理。

一、离群数据的界限

确定离群数据界限的常用方法有如下几种。由于出现离群数据的原因甚多，而这里介绍的一些方法又都是经验性方法，所以，这些方法只能作为实际工作中的参考性方法。

1. 类比法

B.N.斯米尔诺夫根据实际经验，总结出一个确定矿床品位离群数据的界限，见表1-3-3。

表1-3-3 矿床品位离群数据的界限

矿床类型	组分分布性质	典 型 矿 床	离群品位高出平均品位的倍数
Ⅰ	很均匀	大多数沉积矿床	2～3
Ⅱ	均匀	复杂沉积矿床与变质矿床	4～5
Ⅲ	不均匀	绝大多数有色金属矿床	8～10
Ⅳ	很不均匀	大多数稀有金属矿床和金矿床	12～15
Ⅴ	极不均匀	某些稀有金属矿床和金矿床	>15

表1-3-3中的离群品位高出平均品位的倍数是一些经验数据，只能作为参考。从矿床成因角度看，绝大多数的油气藏都属于与沉积岩有关的矿床。所以，确定油气勘探、开发中有关地质问题的离群数据界限时，可以参照表1-3-3中的Ⅰ、Ⅱ矿床类型。

2. 计算法

H.B.沃洛多莫夫给出了下面的公式，通过计算来确定离群数据的界限。

$$ch=c_1+(n-1)c_1M=c_1+\frac{(n-1)c_1(c_1-c_2)}{c_2} \qquad (1\text{-}3\text{-}30)$$

式中 ch——正常数据的最高值，即大于ch的数据则为离群数据；

c_1——校正前（包括离群数据）的样品平均值；

c_2——校正后（不包括离群数据）的样品平均值；

n——包括离群数据在内的样品总数；

$$M=(c_1-c_2)/c_2$$

离群数据在一组数据中，一般只有一、两个，最多也不会超过三个，个数太多时则已不是离群数据。在实际计算时，可令$M=20\sim30\%$，则可计算出离群数据的界限值。

这种方法计算出的ch值显然与子样容量n有关，n越大ch值的偏离可能越大。所以只适合于对小子样的检验。

3. 统计检验法

当一组数据的分布概型属于正态分布，或者这组数据经过某种变换后的分布概型属于正态分布时，可用统计检验法确定离群数据的界限。

（1）小子样的统计检验法 指一组数据的总数n小于等于35个时的统计检验方法。此时，离群数据的界限可用如下方法确定。

$$t_i=\frac{x_i-\bar{x}}{s}, \qquad (i=1,2,\cdots,n) \qquad (1\text{-}3\text{-}31)$$

$$ch=\bar{x}+t_{\alpha}s_{n} \tag{1-3-32}$$

式中 $\bar{x}$——n个数据的平均值；

$$s_{n}=\sqrt{\frac{n-1}{n}}\cdot s$$

s——n个数据的标准差；

x_i——第i个数据值；

t_i——第i个数据的检验值；

ch——离群数据的界限值；

t_{α}——置信水平α下的检验系数，见表1-3-4。

按（1-3-31）或（1-3-32）式计算后，若$t_i>t_{\alpha}$或者$x_i>ch$时，则x_i为离群数据；否则为正常数据。

表1-3-4 t_{α}数值表

n \ α	0.01	0.05	n \ α	0.01	0.05
3	1.414	1.412	15	2.800	2.493
4	1.723	1.689	16	2.837	2.523
5	1.955	1.869	17	2.871	2.551
6	2.130	1.996	18	2.903	2.577
7	2.265	2.093	19	2.932	2.600
8	2.374	2.172	20	2.959	2.623
9	2.464	2.237	21	2.984	2.644
10	2.540	2.294	22	3.008	2.664
11	2.604	2.343	23	3.030	2.683
12	2.663	2.387	24	3.051	2.701
13	2.714	2.426	25	3.071	2.717
14	2.759	2.461			

（2）大子样的统计检验法 指一组数据的总数大于35个时的统计检验方法。

当一组数据的分布概型属于正态分布时，其密度分布为

$$f(x)=\frac{1}{\sqrt{2\pi}\,s}e^{-\frac{(x_i-\bar{x})^2}{2s^2}}\qquad(i=1,2,\cdots,n) \tag{1-3-33}$$

式中 x_i——第i个数据值；

$\bar{x}$——n个数据的平均值；

s——n个数据的标准差。

这种检验方法可称作偏度—峰度检验法。偏度（R_1）也就是偏倚程度，用偏度可以衡量密度分布曲线偏离对称的程度。除一阶中心矩之外，可用任何一个不等0的奇数阶中心矩来衡量密度分布的偏倚程度，通常是采用三阶中心矩作为衡量标准。即

$$R_1=U_1/s^3 \tag{1-3-34}$$

式中　U_1——三阶中心矩，$U_1=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3$。

用峰度(R_2)可以衡量密度分布曲线的陡缓程度，通常是采用四阶中心矩作为衡量标准。即

$$R_2=(U_2/s^4)-3 \tag{1-3-35}$$

式中　U_2——四阶中心矩，$U_2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4$。

可以证明，R_1、R_2近似服从平均值为0，方差为s^*的正态分布，即

$$R_1\sim N(0,s_1^*)$$

$$R_2\sim N(0,s_2^*)$$

其中　$s_1^*=\sqrt{\frac{6}{n}}$，$s_2^*=\sqrt{\frac{24}{n}}$

当给定检验置信水平，例如$\alpha=0.05$时，则有

$$p\left\{-1.96<\frac{R_1}{s_1^*}<1.96\right\}=1-\alpha=0.95$$

$$p\left\{-1.96<\frac{R_2}{s_2^*}<1.96\right\}=1-\alpha=0.95$$

因此，R_1. R_2的否定域可以定为

$$|R_1|>1.96\sqrt{\frac{6}{n}}$$

$$|R_2|>1.96\sqrt{\frac{24}{n}}$$

如果原始数据的离散程度较高，对R_1、R_2的否定域可适当扩大，例如经常采用在原域基础上再加上原域的$\frac{1}{2}$。即

$$p_1=1.96\sqrt{\frac{6}{n}}\left(1+\frac{1}{2}\right)$$

$$p_2=1.96\sqrt{\frac{24}{n}}\left(1+\frac{1}{2}\right)$$

检验时要求

$$|R_1|>p_1 \tag{1-3-36}$$

$$|R_2|>p_2 \tag{1-3-37}$$

这些准备工作完成后，即可对原始数据集合进行密度分布的迭代检验。迭代检验的目的在于确定离群数据的界限值，删除离群数据后，剩下的数据就是由正常数据组成的母体。

实际检验时可分两步进行，第一步是检验数据的密度分布是否服从正态分布；如果不服从正态分布则转入第二步，即删除一些离群数据，一般情况下可用($\bar{x}\pm2s$)作为临界值，凡是大于($\bar{x}+2s$)或小于($\bar{x}-2s$)的数据都要删除，保留剩下的数据。之后，再返回到第一步继续检验保留下来的数据是否服从正态分布，如不服从则转入第二步。如此反复进行这两个步骤，直至保留下来的数据服从正态分布为止。这些数据即可认为是由正常数据组成的

母体。

二、对离群数据的处理

对离群数据的处理要持慎重态度，首先要分析离群数据出现的原因，对于原因不明的离群数据应重新取样观测。

如果经过分析研究后，认为离群数据是错误数据，则必须舍弃。如果不是错误数据，一般可用平均值（全体数据）代替法、邻近数据的平均值代替法、离群数据的界限值代替法、推断值代替法等方法处理离群数据。下面介绍两种最常用的代替离群数据的处理方法。

1. 对一个离群数据的代替方法

当一组数据中只有一个离群数据而又不宜舍弃时，可用（1-3-38）式的计算值代替离群数据：

$$x_p = \frac{x_m + \bar{x}}{2} \qquad (1\text{-}3\text{-}38)$$

式中 x_p——离群数据的代替值；

x_m——离群数据值；

$\bar{x}$——全体数据的平均值。

如果认为x_p值还偏高，也可以用求出的离群数据界限值代替（1-3-38）式中的x_m，而重新计算出x_p值。

2. 对多个离群数据的代替方法

如果一组数据中有r个离群数据，其值为x_{mi}（$i=1,2\cdots,r$），离群数据的平均值为$\bar{x}_m$，此时，可用（1-3-39）式的计算值代替离群数据：

$$x_p = \frac{r-1}{r}\bar{x}_m \qquad (1\text{-}3\text{-}39)$$

当x_p仍然大于离群数据的界限值ch时，也可以用（1-3-40）或（1-3-41）式计算出的x_p代替离群数据：

$$x_p = \frac{r-1}{r}ch \qquad (1\text{-}3\text{-}40)$$

$$x_p = \sqrt{\frac{r-1}{r}}\,ch \qquad (1\text{-}3\text{-}41)$$

第七节 变量的筛选

在地质研究工作中，有时可供使用的地质变量很多。但是，这些变量不一定全是有效变量。为了选出那些有效变量，舍弃那些无效变量，就需要对变量进行逐个筛选。

对变量的筛选往往是多元统计分析成败的重要环节。在以后讲述的多元统计方法中，如逐步回归分析、逐步判别分析、因子分析等方法，都具有筛选变量的功能。而这里介绍的是一种简单易行的筛选变量方法，即0，1化变量筛选法。

一、筛选原理

筛选变量的目的在于从数量较多的M个变量中，选出数量较少的m个有效变量。所谓有效变量是指m个自变量x_1，x_2，…，x_m与因变量y之间的相关关系较好；而x_1，x_2，x_m变量之间的相关关系较差，亦即x_1，x_2，…x_m之间的相互独立性较好。

进行变量筛选时，对于定量数据要变换为0或1，可以首先确定一个门坎值x_k，当$x_i > x_k$时，令$x_i=1$；当$x_i \leqslant x_k$时，令$x_i=0$。因变量y也要变换为0或1。变换后，当因变量y与第i个自变量x_i的值相同时，即$y=1$，$x_i=1$或者$y=0$，$x_i=0$时，可以认为x_i与y之间的相关关系较好，或者说用x_i预报y时预报对了，简称报对；当因变量y与自变量x_i的值不同时，即$y=1$，$x_i=0$或者$y=0$，$x_i=1$时，可以认为x_i与y之间的相关关系较差，或者说用x_i预报y时预报错了，简称报错。按着这种办法可以计算出变量x_i的报对频率n_i。

$$n_i=\frac{N_i}{N} \qquad (i=1,2,\cdots,M) \tag{1-3-42}$$

式中 N_i——用x_i预报y的报对频数；

N——样品总数。

显然，n_i的值越大表示x_i与y之间的相关性越好。但是，只用n_i作为筛选变量的唯一标准是不充分的。这里还需要考虑任何两个自变量之间的相互独立性问题。

如果x_a、x_b分别是变量x_i（$i=1,2,\cdots,M$）中的任意两个变量，现作如下讨论。

（1）当用x_a、x_b预测y，x_a报对，x_b也报对时（即$y=1$，$x_a=1$，$x_b=1$或者$y=0$，$x_a=0$，$x_b=0$），由于两个变量都报对，所以样本中该个体将会增大报对频率。这种情况下，x_a、x_b对于预测y显然是有贡献的。

（2）当x_a报对，而x_b报错时（即$y=1$，$x_a=1$，$x_b=0$或者$y=0$，$x_a=0$，$x_b=1$），虽然x_b报错，但是x_b与x_a的数值正好相反，这说明x_b与x_a之间的相互独立性较好，所以对预测y也能起到较好的作用。

（3）当x_a报错，而x_b报对时（即$y=1$，$x_a=0$，$x_b=1$或者$y=0$，$x_a=1$，$x_b=0$），虽然x_a报错，但是x_a与x_b的数值正好相反，这说明x_a与x_b之间的相互独立性较好，所以对预测y也能起到较好的作用。

（4）当x_a报错，而x_b也报错时，（即$y=1$，$x_a=0$，$x_b=0$或者$y=0$，$x_a=1$，$x_b=1$），由于两个变量都报错，所以样本中该个体将会减小报对频率。这种情况下，x_a、x_b对于预测y显然是有害处的。

从上述4种情况看，前3种情况都有利于预测y。只有最后一种情况，即两个变量同时报错而且两个变量之间又有较好的相关性时，对于预报y必然起到不好作用。因此，就可以用样本中每个变量的报错个体与其他变量同时报错的个体重复出现的次数N_i^*，来衡量变量间相互独立性的好坏，N_i^*可称为变量之间的报错相关频数。由N_i^*可以计算出每个变量与其他变量间的报错相关频率n_i^*'，即

$$n_i^*=\frac{N_i^*}{N(M-1)} \qquad (i=1,2,\cdots,M) \tag{1-3-43}$$

式中 n_i^*——报错相关频率；

N_i^*——报错相关频数；

N——样品总数；

M——变量总数。

综上所述，可以根据每个变量的报对频率n_i及变量间的报错相关频率n_i^*两个指标对变量进行筛选。此处令

$$h_i=\frac{n_i}{n_i^*}=\frac{(M-1)N_i}{N_i^*} \qquad (i=1,2,\cdots,M) \qquad (1\text{-}3\text{-}44)$$

h_i越大，表明自变量x_i与因变量y的相关性越好，并且自变量x_i与其他自变量之间的独立性也越好。反之，h_i越小，表明自变量x_i与因变量y的相关性越差，而且自变量x_i与其他自变量之间的独立性也越差。在筛选变量时，可按h_i值的大小依次剔除不重要的变量，直至剩下的变量都是有效变量为止。最后便可从M个可选用的变量中筛选出m个有效变量。

二、计算过程

为了便于了解计算过程，这里给出一个实际算例。某个天然气勘探地区，由地面地质调查及地震勘探已发现许多局部构造型地质圈闭，为了评价这些构造圈闭的含气性，已收集到了7项评价因素的有关资料。在这众多的构造圈闭中，总共钻探了10个圈闭，其中的5个圈闭已见到工业性气流，而另外5个圈闭钻探落空。这10个圈闭便组成一个已知含气性的子样。为了较准确地评定每个地质变量与圈闭含气性之间的相关性，可用这个子样对7个变量进行筛选。以便于用筛选出的有效变量进一步评价其他未钻地质圈闭的含气性。

已收集到的7项评价地质圈闭含气性的地质变量是：

（1）地质构造的圈闭面积（x_1），面积大于10（km^2）的赋值为1，小于等于10（km^2）的赋值为0；

（2）地质构造的闭合度（x_2），闭合度大于10（m）的赋值为1，小于等于10（m）的赋值为0；

（3）地质构造上有无切顶断层（x_3），无切顶断层时赋值为1，有切顶断层时赋值为0。

（4）地质构造距盆地边界的距离（x_4），距离大于100（km）时赋值为1，小于等于100（km）时赋值为0；

（5）勘探目的层的海拔高度（x_5），在海平面以下的赋值为1，在海平面以上的赋值为0。

（6）构造圈闭的长轴与短轴的长度之比（x_6），比值小于3的赋值为1，比值大于等于3的赋值为0；

（7）构造圈闭的地面裂缝发育程度（x_7），裂缝不发育的赋值为1，裂缝发育的赋值为0。

作为因变量的构造圈闭的含气性（y），钻获工业气流的赋值为1，钻探落空的赋值为0。

这10个地质圈闭的含气性、以及7个地质变量的0，1化变换后的数值详见表1-3-5。此处样品总数$N=10$，可选用的地质变量$M=7$。7个变量预测构造圈闭含气性的报对频率为：

$n_1=0.5$，$n_2=0.4$，$n_3=0.9$，$n_4=0.8$，$n_5=0.9$，$n_6=0.5$，$n_7=0.2$。

可见n_3、n_5最大，即x_3、x_5变量最好，x_4变量次之，而x_7最差。

但是，还要考虑变量之间的独立性，即需要计算变量间的报错相关频数N_i^*。计算结果见表1-3-6。由表中的合计报错相关频数，可以算出每个变量的报错相关频率n_i^*。

$n_1^*=0.18$，$n_2^*=0.23$，$n_3^*=0.02$，$n_4^*=0.10$，$n_5^*=0.02$，$n_6^*=0.20$，$n_7^*=0.25$

表1-3-5 构造圈闭数据表

地质变量 \ 圈闭编号	1	2	3	4	5	6	7	8	9	10
圈闭面积（x_1）	0	1	0	1	1	0	0	1	1	1
闭合度（x_2）	0	1	0	1	0	1	1	0	1	0
有无切顶断层（x_3）	1	1	1	0	1	1	0	1	0	0
与盆地边缘的距离（x_4）	0	1	0	0	0	1	0	1	0	0
目的层海拔高度（x_5）	1	0	0	0	1	1	0	1	0	0
长轴短轴比（x_6）	0	1	0	1	1	1	1	0	1	0
地面裂缝发育程度（x_7）	0	0	1	0	0	0	1	1	1	1
构造圈闭含气性（y）	1	1	0	0	1	1	0	1	0	0

表1-3-6 变量间的报错相关频数

地质变量	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	0	3	0	1	0	3	4
x_2	3	0	0	2	0	5	4
x_3	0	0	0	0	0	0	1
x_4	1	2	0	0	0	1	2
x_5	0	0	0	0	0	0	1
x_6	3	5	0	1	0	0	3
x_7	4	4	1	2	1	3	0
合计	11	14	1	6	1	12	15

最后可以计算出$k_i^{(1)}$值，并进行第一次变量筛选。$k_i^{(1)}$的上角标（1）表示第一次计算结果：

$k_1^{(1)}=2.73$，$k_2^{(1)}=1.71$，$k_3^{(1)}=54.00$，$k_4^{(1)}=8.00$，$k_5^{(1)}=54.00$，$k_6^{(1)}=2.50$，$k_7^{(1)}=0.80$。

可见这7个地质变量中，第7个变量x_7最差，应当首先剔除。

为了进一步筛选变量，应根据表1-3-6计算剩下的6个地质变量的合计报错相关频率n_i^*：

$n_1^*=0.14$，$n_2^*=0.20$，$n_3^*=0$，$n_4^*=0.08$，$n_5^*=0$，$n_6^*=0.18$

再计算$k_i^{(2)}$值，进行第二次变量筛选：

$k_1^{(2)}=3.57$，$k_2^{(2)}=2.00$，$k_3^{(2)}=\infty$，$k_4^{(2)}=10.00$，$k_5^{(2)}=\infty$，$k_6^{(2)}=2.78$。

显然，剩下的6个地质变量中，第2个变量x_2最差，应当剔除。

如果要进行第三次变量筛选，可计算$k_i^{(3)}$：

$k_1^{(3)}=5.00$，$k_3^{(3)}=\infty$，$k_4^{(3)}=16.00$，$k_5^{(3)}=\infty$，$k_6^{(3)}=5.00$，

从计算结果看，x_1及x_6可以同时剔除。最后剩下的x_3、x_4、x_5为3个有效的地质变量。

第八节　取样问题

由于地质问题所涉及的地域、空间十分广阔，因而在研究工作中，往往只能从总体中抽出一部分样品进行研究，即用一个子样来研究总体。但是，所抽取的子样能否代表总体？这就需要研究一下取样问题。

一、随机取样

随机取样要求总体中每个样品被抽到的概率是相等的，每次取样是相互独立的。常用的方法有两种。

1．抽签法

抽签法是最直观的随机取样方法。例如，某个探区经过多年勘探已积累了大量的原油、天然气、地下水的化验室分析数据，每种数据多达几千个，甚至上万个。如果想用随机取样法抽查某种化验项目的100个样品数据时，可将已有的分析数据逐一编上一个号码，并把每个号码写在卡片上，这些卡片经过充分混合后，随机取出100个，那么，卡片上编号对应的100个分析数据就是一个随机子样。

2．随机数被替法

许多数学手册上都附有随机数表，表中的数据之间无任何规律可循，从随机数表的任何一页、任何一行、任何一列的数字开始向上、向下、向左、向右读取位数相同的数字序列就是一组随机数。

如果某项地质数据的数量有N个，想从中随机抽取M个数据（$M \ll N$），组成一个随机子样。可将N个数据逐一进行编号，从1开始编号到N为止。假如N是一个P位数，则可从随机数表中读取M个P位随机数。当随机数的值R大于等于1，并且小于等于N时，可以直接使用；当随机数的值R大于N时，需要将随机数的值减去一个KN，其中$K=1,2,\cdots$，为正整数序列中的某个数，使（$R-KN$）的值在1到N之间。

例如，原有1250个化验分析数据，即$N=1250$，若想从中随机抽取100个数据，组成一个随机子样，即$M=100$。1250是4位数字，因而需要从随机数表中读100个4位随机数。下面列举的随机数是从某个随机数表第一页第5行第11列开始向右读，读完第5行时再从第6行第1列继续读，每个随机数是从所在的行列开始向下读4位，则得到如下随机数序列：

3179，1778，6301，2740，4975，3340，0247，9371，9769，0875，……

这前10个随机数中，只有第7个、第10个分别为247、875，因小于1250可直接使用。其他8个随机数均需要减去1250K（$K=1,2,\cdots$），使其数值在1至1250之间。经过处理后得到：

679，528，51，240，1225，840，247，621，1019，875，……

由这些随机数作为编号的100个数据就是一个随机子样。

实际抽样时，都是由计算机完成。实施步骤是将容量为N的一组数据存入一个数据文件，计算时先将该文件读入一维数组。由介于1到N之间的M个随机数作为下标的全体数据，就是一个容量为M的随机子样。

二、系统取样

系统取样是按着一定的顺序，机械地每隔若干个单位抽取一个样品的方法。例如，在钻井过程中，按钻进深度每增加1m取一包岩屑就是典型的系统取样方法。

这种抽样方法简单易行，但是，有时可能产生系统误差。由于总体的性质不同，被抽样的个体间隔不同，其抽样误差也不相同。

三、分层取样

分层取样是先按某种地质特征把研究对象分为若干个类型、部分或区域等，在统计学上可通称分为若干个层。例如，按不同岩石类型取样，按不同勘探目的层取样，在不同盆地进行取样等等都是分层取样。

分层取样可按随机取样方式进行；如果在各层内抽样比例相同，可称按比例分层抽样。当然，也可以不按比例进行抽样。

分层取样的目的是要把总体分为若干个部分。因此，一般情况下要求每个层的内部差异越小越好，而层与层之间的差异越大越好。

四、群体取样

群体取样不是抽取单个个体，而是抽取由个体组成的若干个集团，即个体群。例如在研究古生态环境时，取单一古生物化石往往不能说明问题，而用由各种古生物个体组成的古生物群落更能反映古生态环境。

群体取样要求群体中包含的个体类型越多越好，以有利于说明其群体特征。

第二篇　地质多元统计分析

地质多元统计分析是运用数理统计方法来研究解决地质研究工作中多变量问题的理论和方法，由于计算机的广泛应用，特别是近年来微型计算机的普及，为地质多元统计分析在地质研究工作中的应用创造了条件。

第一章　回归分析

地质变量之间的关系是比较复杂的，往往是一个地质变量受其他多个地质变量所制约，但是，它们之间又没有确定的函数关系。在这种情况下，常用回归分析方法寻求地质变量之间的统计关系。确切地说，回归分析是研究变量之间相关关系的一种统计分析方法，也就是要建立一个地质变量与另一个地质变量或几个地质变量之间相关关系的数学表达式。

回归分析在地质研究工作中，主要可以解决以下几个方面的问题：

（1）确定一个地质变量与另外一个地质变量或几个地质变量之间是否存在相关关系，如果存在的话，可以找出他们之间相关关系的数学表达式；

（2）根据一个地质变量或几个地质变量的数值，预测另一个地质变量的估计值，并且可以知道这种预测结果的精确度；

（3）如果一个地质变量与另外几个地质变量之间存在相关关系时，可以通过回归分析明确哪些是重要的变量，哪些是次要的变量，哪些是可有可无的变量。

从计算方法上，回归分析可分为一元线性回归分析、一元非线性回归分析、多元线性回归分析、逐步回归分析等主要方法。

第一节　一元线性回归分析

一元线性回归是回归分析中最简单的方法，通过这种分析方法可以得出两个地质变量之间的线性回归模型。

一、一元线性回归方程

如果某一地质变量y与另一个地质变量x之间有相关关系，并且有如下的一元线性回归模型：

$$y=a_0+a_1x+\varepsilon$$

式中的ε是随机误差。若对x与y分别作了n次观测，则可得到n组数据，即

$$y_k,\ x_k,\ e_k \qquad (k=1,2,\cdots,n)$$

假设b_0、b_1是a_0、a_1的估计值，则有如下一元线性回归方程

$$\hat{y}_k = b_0 + b_1 x_k \tag{2-1-1}$$

式中 $\hat{y}_k$——方程给出的回归值，为与观测数据y_k区别起见，可称$\hat{y}_k$为帽y_k；

b_0、b_1——回归方程的待定系数。

现在的问题是如何确定（2-1-1）式中的b_0、b_1，因为b_0、b_1的具体数值目前还不知道，所以可统称为方程的待定系数。为此，需要考察一下实际观测值y_k与回归值$\hat{y}_k$之间的偏差，即

$$G = y_k - \hat{y}_k$$

式中G称为残差（或称为剩余）。残差在回归分析中具有重要的作用。为了确定b_0、b_1，可采用最小二乘法使残差的平方和达到最小，即

$$\min G = \sum_{k=1}^{n}(y_k - \hat{y}_k)^2 = \sum_{k=1}^{n}(y_k - b_0 - b_1 x_k)^2$$

对于实际观测得到的n组数据（y_k, x_k），G是b_0、b_1的非负二次函数，所以必然有极小值存在。因而，G对b_0、b_1的偏导数可满足：

$$\begin{cases} \dfrac{\partial G}{\partial b_0} = 2\sum\limits_{k=1}^{n}(y_k - b_0 - b_1 x_k)(-1) = 0 \\ \dfrac{\partial G}{\partial b_1} = 2\sum\limits_{k=1}^{n}(y_k - b_0 - b_1 x_k)(-x_k) = 0 \end{cases}$$

将该联立方程组展开并移项，可以得到如下联立方程组：

$$\begin{cases} b_0 n + b_1 \sum\limits_{k=1}^{n} x_k = \sum\limits_{k=1}^{n} y_k \\ b_0 \sum\limits_{k=1}^{n} x_k + b_1 \sum\limits_{k=1}^{n} x_k^2 = \sum\limits_{k=1}^{n} x_k y_k \end{cases} \tag{2-1-2}$$

（2-1-2）式写成矩阵形式为：

$$\begin{pmatrix} n & \sum\limits_{k=1}^{n} x_k \\ \sum\limits_{k=1}^{n} x_k & \sum\limits_{k=1}^{n} x_k^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum\limits_{k=1}^{n} y_k \\ \sum\limits_{k=1}^{n} x_k y_k \end{pmatrix} \tag{2-1-3}$$

由（2-1-2）或（2-1-3）式可以解得：

$$b_1 = \frac{\sum\limits_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sum\limits_{k=1}^{n}(x_k - \bar{x})^2} = \frac{\sum\limits_{k=1}^{n} y_k x_k - \dfrac{1}{n}\sum\limits_{k=1}^{n} y_k \sum\limits_{k=1}^{n} x_k}{\sum\limits_{k=1}^{n} x_k^2 - \dfrac{1}{n}\left(\sum\limits_{k=1}^{n} x_k\right)^2} \tag{2-1-4}$$

$$b_0 = \frac{1}{n}\sum_{k=1}^{n} y_k - b_1 \frac{1}{n}\sum_{k=1}^{n} x_k = \bar{y} - b_1 \bar{x} \tag{2-1-5}$$

将求得的b_0、b_1代入（2-1-1）式，便得到了一元线性回归方程。

二、相关系数的显著性检验

前面已经建立了回归方程，但是，这个方程的代表性如何？或者说是否有实际意义？这需要从数学上进行检验，一般称作回归方程的显著性检验。通常用相关系数r_{xy}作为衡量回归方程显著性的标准：

$$r_{xy}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} \tag{2-1-6}$$

式中 $$L_{xx}=\sum_{k=1}^{n}x_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}x_k\right)^2 \tag{2-1-7}$$

$$L_{yy}=\sum_{k=1}^{n}y_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}y_k\right)^2 \tag{2-1-8}$$

$$L_{xy}=\sum_{k=1}^{n}y_kx_k-\frac{1}{n}\sum_{k=1}^{n}x_k\sum_{k=1}^{n}y_k \tag{2-1-9}$$

r_{xy}表示变量x与变量y之间的线性关系的密切程度，$|r_{xy}|\leqslant 1$。

如果比较一下回归方程的系数b_1与相关系数r_{xy}，则有：

$$r_{xy}=\frac{L_{xy}}{\sqrt{L_{xx}L_{yy}}}=\frac{L_{xy}}{L_{xx}}\sqrt{\frac{L_{xx}}{L_{yy}}}$$

$$=b_1\sqrt{\frac{L_{xx}}{L_{yy}}}$$

可见，r_{xy}与b_1有相同的符号。当$r_{xy}>0$时，称x与y正相关；当$r_{xy}<0$时，称x与y负相关。用r_{xy}可以衡量x与y之间的线性相关程度。当$|r_{xy}|=1$时，称为完全线性相关；当$|r_{xy}|=0$时，称为完全线性无关。可见，$|r_{xy}|$越接近1时，线性相关关系越强。

r_{xy}与自由度f及置信度α有关，一元线性回归时的自由度$f=n-2$。当$|r_{xy}|$大于表2-1-1中给出的对应值时，回归方程才认为是显著的，即所建立的回归方程是有代表意义的。

表2-1-1 线性相关系数表

$f=n-2$	$\alpha=0.05$	$\alpha=0.01$	$f=n-2$	$\alpha=0.05$	$\alpha=0.01$	$f=n-2$	$\alpha=0.05$	$\alpha=0.01$
1	0.997	1.000	10	0.576	0.708	19	0.433	0.549
2	0.950	0.990	11	0.553	0.684	20	0.423	0.537
3	0.878	0.959	12	0.532	0.660	21	0.413	0.526
4	0.811	0.917	13	0.514	0.641	22	0.404	0.515
5	0.754	0.874	14	0.497	0.623	23	0.396	0.505
6	0.707	0.834	15	0.482	0.606	24	0.388	0.496
7	0.666	0.798	16	0.468	0.590	25	0.381	0.487
8	0.632	0.765	17	0.456	0.575	26	0.374	0.478
9	0.602	0.735	18	0.444	0.561	27	0.367	0.470

续表

$f=n-2$	$\alpha=0.05$	$\alpha=0.01$	$f=n-2$	$\alpha=0.05$	$\alpha=0.01$	$f=n-2$	$\alpha=0.05$	$\alpha=0.01$
28	0.361	0.463	50	0.273	0.354	125	0.174	0.228
29	0.355	0.456	60	0.250	0.325	150	0.159	0.208
30	0.349	0.449	70	0.232	0.302	200	0.138	0.181
35	0.325	0.418	80	0.217	0.283	300	0.113	0.148
40	0.304	0.393	90	0.205	0.267	400	0.098	0.128
45	0.288	0.372	100	0.195	0.254	1000	0.062	0.081

三、回归方程的F检验

观测值y_k与其平均值$\overline{y}$的离差平方和记为S_{yy}，即

$$S_{yy}=\sum_{k=1}^{n}(y_k-\overline{y})^2=\sum_{k=1}^{n}y_k^2-\frac{1}{n}\left(\sum_{k=1}^{n}y_k\right)^2=L_{yy}$$

S_{yy}可以分解为两部分，即

$$\begin{aligned}S_{yy}&=\sum_{k=1}^{n}(y_k-\overline{y})^2=\sum_{k=1}^{n}[(y_k-\hat{y}_k)+(\hat{y}_k-\overline{y})]^2\\&=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\overline{y})\\&\quad+\sum_{k=1}^{n}(\hat{y}_k-\overline{y})^2\\&=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2+\sum_{k=1}^{n}(\hat{y}_k-\overline{y})^2\\&=V+U\end{aligned}$$

其中，$U=\sum_{k=1}^{n}(\hat{y}_k-\overline{y})^2$，这一部分体现了自变量$x$的变化对因变量$y$的影响，称作回归平方和。另一部分$V=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$，体现了观测值$y_k$与回归值$\hat{y}_k$之间的偏离程度，或者表示了误差的大小，称作剩余平方和。

回归平方和U的自由度$f_U=1$，剩余平方和V的自由度$f_V=n-2$，总的离差平方和S_{yy}的自由度$f_{yy}=n-1$。

如果自变量x与因变量y之间有线性关系时，则可建立统计量F

$$F=\frac{U/1}{V/(n-2)}\sim F_\alpha(1,n-2) \qquad (2\text{-}1\text{-}10)$$

式中　$F_\alpha(1,n-2)$——在置信水平α下，第1自由度为1，第2自由度为$n-2$的F分布。

在给定置信水平α下，回归方程的F值应大于F_α值，α一般取0.1，0.05，0.01，而（1-α）表示F检验的可靠程度。

当$F>F_a$时，认为x与y有明显的线性关系；当$F\leqslant F_a$时，认为x与y没有明显的线性关系，即所建立的回归方程无意义。

四、残差分析与回归预报

回归方程通过相关系数的显著性检验或F检验就能确定所建立的方程是否有代表性；如果方程是有实际意义的，但是，并不能认为n组观测数据都是可信的。那么，如何判断数据中是否有离群数据呢？这就是残差分析所要解决的问题。

所谓残差就是观测值y_k与回归值$\hat{y}_k$之间的偏差，即$v_k=y_k-\hat{y}_k(k=1,2,\cdots,n)$。利用$v_k$分析观测数据的可靠性就是残差分析。

$$V=\sum_{k=1}^{n}v_k^2=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

剩余平方和v除以它的自由度f_e所得的商

$$s^2=\frac{v}{f_e}=\frac{v}{n-2}=\frac{1}{n-2}\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

称为剩余方差，s^2可以作为在排除了x对y的线性影响后，衡量y随机波动大小的一个估计值。而剩余方差的方根，即

$$s=\sqrt{\frac{v}{n-2}}=\sqrt{\frac{1}{n-2}\sum_{k=1}^{n}(y_k-\hat{y}_k)^2} \qquad (2\text{-}1\text{-}11)$$

称为剩余标准差。s可以用来衡量所有随机因素对y的平均偏差的大小。

对于服从正态分布的自变量x，当x等于某个确定值x_0时，因变量y的取值也服从正态分布，y的平均值$\overline{y}$就是$x=x_0$时，回归方程

$$\hat{y}_0=b_0+b_1x_0$$

的回归值。y的取值以$\hat{y}_0$为中心呈对称分布，越靠近$\hat{y}_0$的地方出现的机率越大，而越远离$\hat{y}_0$的地方出现的机率越小，并且与剩余标准差有如下关系：

落在$(\hat{y}_0-s,\ \hat{y}_0+s)$区间内的$y$值约占68%

落在$(\hat{y}_0-2s,\hat{y}_0+2s)$区间内的$y$值约占95%

落在$(\hat{y}_0-3s,\hat{y}_0+3s)$区间内的$y$值约占99.7%

这种关系对于实际研究工作是有意义的，s越小，y的取值范围就越小。因此，通常在回归方程$\hat{y}=b_0+b_1x$的两侧作两条与回归方程平行的直线方程：

$$y'=b_0+b_1x-2s \qquad (2\text{-}1\text{-}12)$$

$$y''=b_0+b_1x+2s \qquad (2\text{-}1\text{-}13)$$

作为y值出现的区间，这就是通常所说的两倍剩余标准差预测准则。对应于自变量x的因变量y的预测值，出现在这两条直线范围内的机率为95%。而对于出现在这两条直线范围以外的数据要进行检验，判断其是否为离群数据。

五、加权回归分析

在实际的地质研究工作中，有时已得到的观测数据在所讨论的问题中，各个数据的重要性并不相同，也就是说有些数据比较重要，而有些数据比较次要。对这样的数据进行回归分析时就不能平等看待，而要对重要的数据赋以较大的权，对次要的数据赋以较小的权。用最小二乘法确定回归方程的待定系数时，使得残差平方的加权和达到最小，即

$$\min G=\sum_{k=1}^{n}t_k(y_k-\hat{y}_k)^2=\sum_{k=1}^{n}t_k(y_k-b_0-b_1x_k)^2$$

求G对b_0, b_1的偏导数，并令其为0则有

$$\begin{cases} b_0\sum_{k=1}^{n}t_k+b_1\sum_{k=1}^{n}t_kx_k=\sum_{k=1}^{n}t_ky_k \\ b_0\sum_{k=1}^{n}t_kx_k+b_1\sum_{k=1}^{n}t_kx_k^2=\sum_{k=1}^{n}t_kx_ky_k \end{cases} \tag{2-1-14}$$

自变量x，因变量y的平均值可分别记为

$$\overline{x}_t=\sum_{k=1}^{n}t_kx_k\Big/\sum_{k=1}^{n}t_k \tag{2-1-15}$$

$$\overline{y}_t=\sum_{k=1}^{n}t_ky_k\Big/\sum_{k=1}^{n}t_k \tag{2-1-16}$$

由（2-1-14）、（2-1-15）、（2-1-16）式可得

$$b_0=\overline{y}_t-b_1\overline{x}_t \tag{2-1-17}$$

$$b_1=\frac{\sum_{k=1}^{n}t_kx_ky_k-\sum_{k=1}^{n}t_kx_k\overline{y}_t}{\sum_{k=1}^{n}t_kx_k^2-\sum_{k=1}^{n}t_kx_k\overline{x}_t}$$

$$=\frac{\sum_{k=1}^{n}t_k(y_k-\overline{y}_t)(x_k-\overline{x}_t)}{\sum_{k=1}^{n}t_k(x_k-\overline{x}_t)^2} \tag{2-1-18}$$

加权回归剩余标准差为

$$s_t=\sqrt{\frac{1}{n-2}\sum_{k=1}^{n}t_k(y_k-\hat{y}_k)^2} \tag{2-1-19}$$

六、算　　例

［1］ 苏联的И.И.涅斯乔洛夫等人，根据世界上22个勘探程度较高的含油气盆地资料，用一元线性回归分析方法得到，沉积盆地的油气总地质储量与该盆地的平均沉积速度之

间成对数线性函数关系，即

$$\lg Q=2.183+1.613\lg V$$

式中　Q——沉积盆地的油气总地质储量（10^8t）；

　　　V——盆地的平均沉积速度（kkm^3/Ma）。

这个一元线性回归方程的原函数是个幂函数：

$$Q=aV^b$$

如果令　$y=\lg Q, x=\lg V$则有

$$y=\lg a+ax$$

即　$\lg a=2.183$，$b=1.613$。

［2］ 镜质体反射率是划分生油岩成熟度的有效指标。松辽盆地南部61个样品的镜质体反射率R_o与时间—温度指数（TTI）之间的相关关系较好。其一元线性回归方程为

$$R_o=0.219+0.493\lg TTI$$

相关系数$r=0.99$，线性回归方程的图形见图2-1-1。

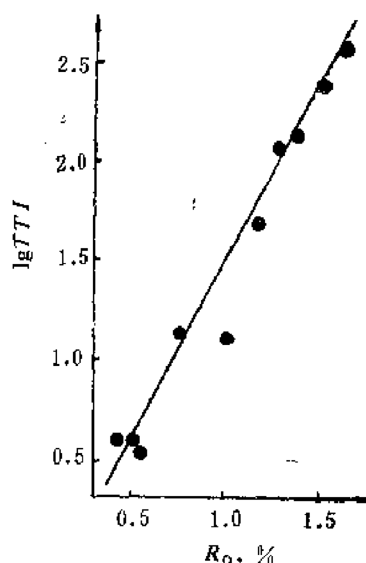


图2-1-1　R_o与$\lg TTI$之间的线性关系图

［3］ 朱学愚根据河南新乡地区忠义县试验厂1954～1964年的地下水位动态观测资料，找出每年雨季（6～8月）时，每次降水量和地下水位升高幅度之间的相关关系。资料是从地下水动态观测的实际资料中选取的，每次降水量小于20mm者不计在内，见表2-1-2。

令降水量为自变量x，地下水位的变幅为因变量y。经计算得到如下回归方程：

$$y-0.67=0.0069(x-107.13)$$

回归方程的相关系数$r=0.89$。

用该方程对1965年9月4日降水后水位升高幅度y作了预报，9月4日降水量为$x=39.5$mm，回归方程预报的地下水位升高幅度为

表2-1-2　降水量与地下水位升高的观测数据

序号	时间（年，月，日）	降水量 x（mm）	地下水位变幅 y（m）	序号	时间（年，月，日）	降水量 x（mm）	地下水位变幅 y（m）
1	54,5,13	45.8	0.25	11	57,7,22	70.2	0.43
2	54,7,10～12	61.4	0.43	12	58,7,5～9	141.8	0.85
3	54,8,3～9	172.7	1.09	13	58,8,2	82.7	0.42
4	54,8,18～19	50.2	0.12	14	63,8,6～9	117.6	0.71
5	55,7,1～8	148.8	0.28	15	63,8,20～21	107.7	0.66
6	56,6,2～3	71.9	0.24	16	64,7,16～17	92.7	0.69
7	56,6,18～24	212.2	1.48	17	64,7,25～27	88.7	0.84
8	56,7,28～30	170.1	1.58	18	64,8,11～12	64.9	0.64
9	56,8,25	35.0	0.25	19	64,8,17	41.7	0.27
10	57,7,10～15	274.2	1.85	20	64,8,28	92.6	0.55

$$y=0.0069(39.5-107.13)+0.67=0.202m$$

而地下水位升高幅度的实测值为0.23m。这表明回归方程的预报值与实测值基本相符。可见这一回归方程可用于预报雨季降水时的地下水位升高幅度，有利于掌握地下水位的动态变化。

第二节 多元线性回归分析

多元线性回归分析是适用于一个变量与多个变量之间具有线性相关关系时的统计分析方法，通过这种分析方法可以建立它们之间的线性回归方程。

一、多元线性回归方程

如果某一地质变量y与m个地质变量x_i（$i=1, 2, \cdots, m$）之间具有相关关系，并且有如下m元线性回归模型

$$y=a_0+a_1x_1+a_2x_2+\cdots+a_mx_m+\varepsilon$$

式中的ε是随机误差。若对y及x_i分别作了n次观测，则有n组数据，即

$$y_k,\ x_{ik},\ \varepsilon_k\ (k=1, 2, \cdots, n;\ i=1, 2, \cdots, m)$$

其中的ε_k是服从正态分布$N(0, \sigma_\varepsilon^2)$的$n$个相互独立的随机变量。

假设b_0，b_1，b_2，…，b_m是a_0，a_1，a_2，…a_m的估计值，则有如下回归方程：

$$\hat{y}_k=b_0+b_1x_{1k}+b_2x_{2k}+\cdots+b_mx_{mk} \qquad (2\text{-}1\text{-}20)$$

式中 $\hat{y}_k$——回归方程给出的回归值，为与观测值y_k区别起见，可称$\hat{y}_k$为帽y_k。

实际观测值y_k与回归值$\hat{y}_k$的差，即

$$G=y_k-\hat{y}_k\ (k=1, 2, \cdots, n)$$

称为残差（或称为剩余），残差表示了观测值与回归值之间的偏差。

现在的问题是要确定（2-1-20）式中的待定系数b_0，b_1，b_2，…，b_m，这里可采用最小二乘法使残差平方和达到最小，即

$$\min G=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2=\sum_{k=1}^{n}(y_k-b_0-b_1x_1-b_2x_2-\cdots-b_mx_m)^2$$

对于观测得到的n组数据（y_k，x_{ik}），G是$b_0, b_1, b_2\cdots, b_m$的非负二次函数，因而一定有极小值存在。所以，b_0，b_1，b_2，…，b_m必然满足

$$\begin{cases}\dfrac{\partial G}{\partial b_0}=2\sum\limits_{k=1}^{n}(y_k-b_0-b_1x_{1k}-b_2x_{2k}-\cdots-b_mx_{mk})(-1)=0\\[2ex]\dfrac{\partial G}{\partial b_i}=2\sum\limits_{k=1}^{n}(y_k-b_0-b_1x_{1k}-b_2x_{2k}-\cdots-b_mx_{mk})(-x_{ik})=0\end{cases}$$

$$(i=1, 2, \cdots, m)$$

这是$m+1$个待定系数b_0，b_1，b_2，…，b_m的正规联立方程组。由第一个方程可以得到：

$$b_0=\bar{y}-\sum_{j=1}^{m} b_j\bar{x}_j \tag{2-1-21}$$

式中 $\bar{x}_j=\frac{1}{n}\sum_{k=1}^{n} x_{jk}$； $\bar{y}=\frac{1}{n}\sum_{k=1}^{n} y_k$

$(j=1,\ 2,\ \cdots,\ m)$

将b_0代入上面正规联立方程组中的其他各方程中，则有

$$0=\sum_{k=1}^{n}(y_k-b_0-b_1x_{1k}-b_2x_{2k}-\cdots-b_mx_{mk})(x_{jk})$$

$$=\sum_{k=1}^{n}\left(y_k-\bar{y}+\sum_{i=1}^{m}b_i\bar{x}_i-\sum_{i=1}^{m}b_ix_{ik}\right)(x_{jk}-\bar{x}_j)$$

$$=\sum_{k=1}^{n}(y_k-\bar{y})(x_{jk}-\bar{x}_j)-\sum_{k=1}^{n}\sum_{i=1}^{m}b_i(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$$

$$=\sum_{k=1}^{n}(x_{jk}-\bar{x}_j)(y_k-\bar{y})-\sum_{i=1}^{m}b_i\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)$$

$$=S_{jy}-\sum_{i=1}^{m}b_iS_{ij} \qquad (j=1,\ 2,\ \cdots,\ m)$$

因而有 $\sum_{i=1}^{m}b_iS_{ij}=S_{jy}$

写成展开形式为

$$\begin{cases}S_{11}b_1+S_{12}b_2+\cdots+S_{1m}b_m=S_{1y}\\S_{21}b_1+S_{22}b_2+\cdots+S_{2m}b_m=S_{2y}\\\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\\S_{m1}b_1+S_{m2}b_2+\cdots+S_{mm}b_m=S_{my}\end{cases} \tag{2-1-22}$$

式中 $S_{ij}=\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)\quad(i,j=1,2,\cdots,\ m)$

$$S_{iy}=\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(y_k-\bar{y})\qquad(i=1,2,\cdots,m)$$

由方程组（2-1-22）式可以解出待定系数b_1，b_2，…，b_m，再加上由（2-1-21）式求出的b_0，回代到（2-1-20）式中，即可得到多元线性回归方程。

为了便于求解回归方程的待定系数，可以把方程组写成矩阵形式，采用逆矩阵求解待定系数。令

$$C=\begin{pmatrix}x_{11}-\bar{x}_1 & x_{21}-\bar{x}_2\cdots x_{m1}-\bar{x}_m\\x_{12}-\bar{x}_1 & x_{22}-\bar{x}_2\cdots x_{m2}-\bar{x}_m\\\cdots\quad\cdots & \cdots\quad\cdots\quad\cdots\\x_{1n}-\bar{x}_1 & x_{2n}-\bar{x}_2\cdots x_{mn}-\bar{x}_m\end{pmatrix}$$

$$b=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

$$Y=\begin{pmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}$$

那么，（2-1-22）式可写成

$$C'Cb=C'Y$$

如果记$C'C=S$，$C'Y=B$

则有 $Sb=B$ （2-1-23）

当系数矩阵满秩条件下（该条件一般容易满足），待定系数的解可用矩阵表示为

$$b=S^{-1}B=(C'C)^{-1}C'Y \qquad (2\text{-}1\text{-}24)$$

二、回归方程及回归系数的显著性检验

前面建立的回归方程是由已知的n次观测值，用最小二乘法求出待定系数的。但是，这个回归方程的代表性如何？各个地质变量在回归方程中的作用如何（也称贡献如何）？这需要从数学上进行检验，这就是回归方程的显著性检验。

令观测值y_k与其平均值$\bar{y}$之间的离差平方和为S_{yy}，即

$$S_{yy}=\sum_{k=1}^{n}(y_k-\bar{y})^2$$

S_{yy}可以分解为两部分

$$\begin{aligned} S_{yy}&=\sum_{k=1}^{n}(y_k-\bar{y})^2=\sum_{k=1}^{n}[(y_k-\hat{y}_k)+(\hat{y}_k-\bar{y})]^2 \\ &=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2+2\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{y}) \\ &+\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2 \\ &=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2+\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2 \end{aligned}$$

其中的交叉项等于0，证明如下。由b_0、b_i对G的偏导数方程可以得到

$$\begin{cases} \sum_{k=1}^{n}(y_k-\hat{y}_k)=0 \\ \sum_{k=1}^{n}(y_k-\hat{y}_k)x_{ik}=0 \end{cases}$$

由（2-1-21）式知道

$$\bar{y}=b_0+b_1\bar{x}_1+b_2\bar{x}_2+\cdots+b_m\bar{x}_m$$

同时有

$$\hat{y}_k-\bar{y}=b_1(x_{1k}-\bar{x}_1)+b_2(x_{2k}-\bar{x}_2)+\cdots+b_m(x_{mk}-\bar{x}_m)$$

$$=\sum_{i=1}^{m}b_i(x_{ik}-\bar{x}_i)$$

最后有

$$\sum_{k=1}^{n}(y_k-\hat{y}_k)(\hat{y}_k-\bar{y})=\sum_{k=1}^{n}(y_k-\hat{y}_k)\sum_{i=1}^{m}b_i(x_{ik}-\bar{x}_i)$$

$$=\sum_{i=1}^{m}b_i\sum_{k=1}^{n}(y_k-\hat{y}_k)(x_{ik}-\bar{x}_i)$$

$$=\sum_{i=1}^{m}b_i\Big[\sum_{k=1}^{n}(y_k-\hat{y}_k)x_{ik}-\bar{x}_i\sum_{k=1}^{n}(y_k-\hat{y}_k)\Big]$$

$$=0$$

因而证明了交叉项等于0，也就是说离差平方和S_{yy}是由两部分组成。其中的一部分为

$$\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2=\sum_{k=1}^{n}[b_0+b_1x_{1k}+b_2x_{2k}+\cdots+b_mx_{mk}-\bar{y}]^2$$

$$=\sum_{k=1}^{n}[b_1(x_{1k}-\bar{x}_1)+b_2(x_{2k}-\bar{x}_2)+\cdots+b_m(x_{mk}-\bar{x}_m)]^2$$

这一部分表现了自变量x_1，$x_2,\cdots,x_m$的变化对y的影响，称为回归平方和，记为U，即

$$U=\sum_{k=1}^{n}(\hat{y}_k-\bar{y})^2$$

其自由度为$f_U=m$。

另一部分表现了观测值y_k与回归值$\hat{y}_k$之间的离差平方和，它表现了误差的影响，称为剩余平方和，记为V，即

$$V=\sum_{k=1}^{n}(y_k-\hat{y}_k)^2$$

其自由度为$f_V=n-m-1$。

这两个部分合起来，即

$$S_{yy}=V+U$$

总的自由度$f=f_V+f_U=n-1$。

需要指出，在我们研究变量y与变量$x_1,x_2,\cdots,x_m$之间的相关关系时，事先并不知道它们之间是否存在线性关系，因此要对回归方程进行检验。如果线性回归方程无显著意义，则

地质变量x_i（$i=1, 2, \cdots, m$）前面的待定系数都可以取值为0；而如果方程有显著意义就不能取值为0。所以，检验线性回归方程是否有意义等价于检验如下假设，亦即

$$H_0: a_1 = a_2 = \cdots = a_m = 0$$

为此，可建立统计量F

$$F = \frac{\dfrac{U}{f_U}}{\dfrac{V}{f_V}} = \frac{\dfrac{U}{m}}{\dfrac{V}{n-m-1}} \qquad (2\text{-}1\text{-}25)$$

可以证明，若H_0成立则统计量F服从$F(m, n-m-1)$分布，因而对于给定检验水平α，可由F分布数据表查到F_α的数值，如果统计量$F > F_\alpha$，就在检验水平α下，拒绝原来假设H_0，即认为线性回归方程有显著意义。反之，如果统计量$F < F_\alpha$，就在检验水平α下，接受原来假设H_0，即认为线性回归方程无显著意义。

在进行方程显著性检验时，还要检验每个变量在方程中的作用。一般总希望在所建立的方程中去掉那些可有可无的次要变量，使方程中仅保留一些重要的变量。显然，如果某个变量x_i对变量y的作用不显著，则相当于x_i的待定系数b_i可取值为0。因此，检验第i个变量x_i的显著性等价于检验假设

$$H_0: a_i = 0$$

为此，可建立统计量F_i

$$F_i = \frac{\dfrac{(b_i - a_i)^2}{S_{ii}^{-1}}}{\dfrac{V}{n-m-1}}$$

由于假设$a_i = 0$，所以其原假设H_0可采用

$$F_i = \frac{\dfrac{b_i^2}{S_{ii}^{-1}}}{\dfrac{V}{n-m-1}} \qquad (2\text{-}1\text{-}26)$$

来检验回归系数a_i的显著性。（2-1-26）式中的S_{ii}^{-1}是正规方程组系数矩阵S的逆矩阵S^{-1}中的第i行第i列的元素。

对于给定检验水平α，查F分布表可以得到$F_\alpha(1, n-m-1)$，如果统计量$F_i > F_\alpha$，则在检验水平α下，拒绝假设$H_0: a_i = 0$，即认为变量x_i应保留在方程中。反之，如果统计量$F_i < F$，则在检验水平α下，接受原假设$H_0: a_i = 0$，即认为变量x_i应从方程中去掉。

当对m个变量x_i（$i=1, 2, \cdots, m$）逐个进行检验，并且去掉了不显著的变量后，则可以对留下来的变量重新建立更为简单而有效的线性回归方程。

如果经过检验，有p个变量x_i（$i=1, 2, \cdots, P$）可以留在方程中，则有

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

利用这个多元线性回归方程就可以进行回归预报与质量控制。

所谓回归预报，是指对于新给定的一组自变量数值$x_1', x_2', \cdots, x_p'$，如何估计出其所对应的y'值问题。当然，最直接的办法就是将$x_1', x_2', \cdots, x_p'$代入回归方程来计算y'，即

$$\hat{y}'=b_0+b_1x_1'+b_2x_2'+\cdots+b_px_p'$$

用$\hat{y}'$作为y'的估计值。y'落在$(\hat{y}'-s,\ \hat{y}'+s)$区间内的概率为68%；落在$(\hat{y}'-2s,\hat{y}'+2s)$区间内的概率为95%；落在$(\hat{y}'-3s,\hat{y}'+3s)$区间内的概率为99.7%。这里的$s$为剩余标准差，即

$$s=\sqrt{\frac{V}{n-p-1}}$$

所谓质量控制问题，其实质是回归预报的反问题。质量控制在工业生产中有广泛的用途，但在地质研究工作中并不多见。它所要研究的问题是如果要求y'出现在指定的范围$[y_1',y_2']$内，应当如何控制x_1'，x_2'，…，x_p'的取值质量。这里要求所取的x_1'，x_2'，…，x_p'应满足如下条件

$$\hat{y}'-2s\geqslant y_1';\quad \hat{y}'+2s\leqslant y_2'$$

其中$\hat{y}'$是x_1',x_2'，…，x_p'的回归值，而s是剩余标准差。

三、算　例

[1]　我国东部地区一些含油凹陷的实际资料表明，凹陷中单位面积内的石油资源量（或石油储量）与以下5个地质变量之间有相关关系，其多元线性回归方程为

$$Q=b_0+b_1\frac{V_{生}}{V_{沉}}+b_2\frac{H}{C}+b_3\frac{V_{储}}{V_{沉}}+b_4\frac{S_{近}}{S_{沉}}+b_5N$$

式中　Q——凹陷中单位面积的石油资源量（或石油储量）；

$V_{生}$——生油岩体积；

$V_{沉}$——沉积岩体积；

H——总烃含量；

C——有机碳含量；

$V_{储}$——储集层体积；

$S_{近}$——近油源圈闭面积；

$S_{沉}$——沉积岩面积；

N——剥蚀次数。

b_0，b_1，b_2，…，b_5是回归方程的待定系数，对于不同地区，这些系数的值是不相同的。

朱子仁等根据中原油田的实际资料建立的多元线性回归方程为

$$Q=-6.654+0.835\frac{V_{生}}{V_{沉}}+0.597\frac{H}{C}+0.269\frac{V_{储}}{V_{沉}}+0.142\frac{S_{近}}{S_{沉}}-0.05N$$

式中　Q——凹陷中单位面积的石油资源量（$10^4t/km^2$）。

根据这一回归方程，预测前梨园洼陷沙二下段460km^2面积内的石油资源总量ΣQ为

$$Q=28.834(10^4t/km^2)$$

$$\Sigma Q=28.834\times460=13263.64(10^4t)$$

[2]　W.C.克鲁滨（1970）在美国肯塔基州东部选择了一个地质条件上相对比较均匀

的地区，这个地区有许多大小不等的排水盆地，他选用了所有的三级盆地。并在这些盆地中测量了以下7个地质变量：

①盆地排水口高度x_1(ft❶)；

②盆地的起伏x_2(ft)；

③盆地面积x_3(km^2)；

④盆地中河流总长度x_4(m)；

⑤水系密度，即盆地内河流总长度除以盆地的面积x_5(m/km^2)；

⑥盆地形态，即盆地内接圆与外接圆的比值x_6；

⑦盆地的大小，即源流的数目y。

这7个变量中的前6个变量作为自变量，最后一个变量作为因变量。进行多元线性回归分析的目的是确定自变量x_1，x_2，…，x_6对因变量y的影响。在研究区内，总共测量了92个三级盆地的变量值，计算时实际使用了50个盆地的资料，见表2-1-3。

经计算得到如下多元线性回归方程：

$$y=-2.24+0.01x_1+0.02x_2-23.28x_3+6.26x_4-0.20x_5-11.66x_6$$

为了检验方程的显著性，计算了F值

$$F=\frac{U/f_U}{V/f_V}=\frac{1800.70/6}{1134.12/43}=11.38$$

查表得到$F_{0.05}(6,43)=2.34$，而

$$F=11.38>F_{0.05}=2.34$$

这说明所建立的回归方程是有意义的。也就是说，用盆地排水口高度、盆地的起伏、盆地的面积、盆地中河流总长度、水系密度、盆地形态6个变量可以对盆地的大小作出预报。

表2-1-3　肯塔基州东部三级盆地的7个地貌变量值

序号	y	x_1 (ft)	x_2 (ft)	x_3 (km^2)	x_4 (m)	x_5 (m/km^2)	x_6
1	14	720	570	0.07	154	2200	61
2	6	670	610	0.03	80	2667	62
3	5	860	550	0.11	84	763	62
4	7	870	610	0.11	122	1110	63
5	11	730	570	0.14	185	1321	52
6	14	690	590	0.12	200	1667	50
7	12	880	640	0.11	170	1545	41
8	18	760	690	0.28	340	1215	57
9	6	820	600	0.05	100	2000	41
10	5	720	480	0.03	80	2667	60
11	17	670	670	0.19	290	1526	51
12	5	660	600	0.05	90	1800	53
13	22	830	660	0.18	260	1444	57
14	7	780	620	0.17	111	652	57

❶ 1ft=0.3048m

续表

序号	y	x_1 (ft)	x_2 (ft)	x_3 (km²)	x_4 (m)	x_5 (m/km²)	x_6
15	16	750	740	0.15	184	1227	67
16	17	770	630	0.21	227	1080	59
17	6	750	570	0.04	60	1500	65
18	18	750	580	0.20	259	1295	39
19	14	740	760	0.09	62	689	64
20	21	750	740	0.06	96	1582	53
21	22	750	760	0.11	105	954	64
22	23	740	770	0.32	350	1094	55
23	28	940	510	0.21	232	1105	52
24	42	700	600	0.23	266	1156	34
25	22	810	580	0.44	396	886	29
26	10	920	500	0.13	142	1097	65
27	11	920	490	0.12	145	1208	72
28	12	790	605	0.33	253	766	59
29	13	860	550	0.23	241	1048	78
30	31	860	630	0.87	702	807	55
31	18	880	520	0.37	288	778	51
32	13	730	460	0.17	162	958	40
33	4	720	440	0.08	67	838	60
34	5	780	300	0.03	52	1738	57
35	9	700	460	0.10	121	1210	50
36	13	680	520	0.26	220	846	41
37	10	820	520	0.08	123	1537	51
38	13	710	520	0.24	238	992	41
39	13	800	440	0.19	231	1216	51
40	11	700	510	0.16	178	1113	76
41	12	675	570	0.18	169	933	42
42	4	740	510	0.08	65	812	49
43	17	740	520	0.31	334	1078	67
44	9	770	600	0.21	184	876	47
45	8	820	520	0.11	136	1237	56
46	13	850	490	0.22	233	1059	74
47	22	820	629	0.34	410	1206	39
48	10	820	610	0.11	149	1364	60
49	19	680	640	0.46	348	757	55
50	27	660	789	0.55	387	695	38

第三节　逐步回归分析

逐步（线性）回归分析是在多元线性回归分析基础上衍生出来的一种技巧性算法。这种算法的优点是可以从数量较多的变量中，能够自动筛选出最重要的变量进入回归方程，从而避免了多元线性回归分析中，要逐个对所有变量进行显著性检验，以及再重新建立回归方程的繁锁步骤。

逐步回归分析的要点是在计算过程中，根据自变量$x_i(i=1, 2, \cdots, m)$对因变量y的重要性，依次引进到方程中，同时还要对已引进的变量逐个检验，通过检验保留有用的变量，剔除无用的变量。如此办理，边引进边剔除，直到既不能引进也不能剔除时为止。

一、变量的"引进"与"剔除"

1. "引进"变量的准则

假如原有的m个地质变量，已引进p个变量进入了方程，即有如下回归方程

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$$

方程的离差平方和分解公式是

$$S_{yy}=U+V$$

为表示U、V与已引进变量的关系，可用符号$U(x_1, x_2, \cdots, x_p)$、$V(x_1, x_2, \cdots, x_p)$表示。

如果再引进一个变量$x_i(i=p+1, p+2, \cdots, m)$，则新方程相应的离差平方和分解公式是

$$S_{yy}=U(x_1, x_2, \cdots, x_p, x_i)+V(x_1, x_2, \cdots, x_p, x_i)$$

而与未引进x_i之前的离差平方和分解公式

$$S_{yy}=U(x_1, x_2, \cdots, x_p)+V(x_1, x_2, \cdots, x_p)$$

相比较，两式的左端是一样的，即

$$S_{yy}=\sum_{k=1}^{n}(y_k-\bar{y})^2$$

这说明离差平方和S_{yy}与变量x_i的进入无关。但是，引进一个新的变量x_i后，会使回归平方和从$U(x_1, x_2\cdots, x_p)$增加到$U(x_1, x_2, \cdots, x_p, x_i)$；而残差平方和则从$V(x_1, x_2, \cdots, x_p)$降低到$V(x_1, x_2, \cdots, x_p, x_i)$。并且有

$$U(x_1, x_2, \cdots, x_p, x_i)-U(x_1, x_2, \cdots, x_p)=\Delta U$$
$$=V(x_1, x_2, \cdots, x_p)-V(x_1, x_2, \cdots, x_p, x_i)=\Delta V$$

令：$W_i=\Delta U$

那么，W_i可看作是新引进变量x_i对回归平方和的"贡献"，或者说W_i是x_i对y的方差贡献。这种贡献可以理解为x_i进入方程给地质变量y带来的信息。

因而，可将这个W_i与剩余平方和V进行比较，看看x_i的影响是否显著，即可用统计量$F_{1,i}$对x_i进行检验。

$$F_{1i}=\frac{W_i/1}{V(x_1,x_2,\cdots,x_p,x_i)/(n-p-2)} \tag{2-1-27}$$

式中 p——已引进回归方程中的变量个数；

n——子样容量。

如果选定合适的临界值$F_{进}$，而当$F_{1i}>F_{进}$时，表明新引进的变量x_i是有意义的，因此应当进入方程。而当$F_{1i}\leqslant F_{进}$时，表明这一变量x_i是无意义的，因而不能进入方程。

当然，满足$F_{1i}>F_{进}$的变量可能不止一个，此时可选贡献最大的变量首先进入方程。即选

$$\max_{p<i\leqslant m} F_{1i}$$

F_{1i}值为最大的变量首先进入回归方程。

2. "剔除"变量的准则

假如有p个变量已引进到方程中，即有

$$\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_px_p$$

相应的离差平方和分解公式为

$$S_{yy}=U(x_1,\ x_2,\ \cdots,\ x_p)+V(x_1,\ x_2,\ \cdots,\ x_p)$$

如果从已引进的p个变量中，剔除$x_i(i=1,\ 2,\ \cdots,\ p)$后，相应的离差平方和分解公式为

$$\begin{aligned}S_{yy}=&U(x_1,\ x_2,\ \cdots,\ x_{i-1},\ x_{i+1},\ \cdots,\ x_p)\\&+V(x_1,\ x_2,\ \cdots,\ x_{i-1},\ x_{i+1},\ \cdots,\ x_p)\end{aligned}$$

与前面离差平方和分解公式相比，回归平方和从$U(x_1,\ x_2,\ \cdots,\ x_p)$降低到$U(x_1,\ x_2,\cdots,\ x_{i-1},\ x_{i+1},\ \cdots,\ x_p)$；而残差平方和从$V(x_1,\ x_2,\ \cdots,\ x_p)$增加到$U(x_1,\ x_2,\ \cdots,\ x_{i-1},\ x_{i+1},\ \cdots,\ x_p)$，并且有

$$\begin{aligned}&U(x_1,\ x_2,\ \cdots,\ x_p)-U(x_1,\ x_2,\ \cdots,\ x_{i-1},\ x_{i-2},\ \cdots,\ x_p)=\Delta U\\=&V(x_1,\ x_2,\ \cdots,\ x_{i-1},\ x_{i+1},\ \cdots,\ x_p)-V(x_1,\ x_2,\ \cdots,\ x_p)=\Delta V\end{aligned}$$

那么，被剔除的变量x_i对方程减少的贡献是

$$W_i=\Delta U$$

同样，可用统计量F_{2i}进行检验

$$F_{2i}=\frac{W_i/1}{V(x_1,\ x_2,\ \cdots,\ x_p)/(n-p-1)} \tag{2-1-28}$$

F_{2i}越小表示x_i留在方程中的意义越小。此时应当选p个变量中贡献最小的变量首先从方程中剔除。即选

$$\min_{1\leqslant i\leqslant p} F_{2i}$$

F_{2i}值为最小的变量首先从方程中剔除。

3. 矩阵变换方法

为了便于讨论问题，需要对前一节多元线性回归方程中的（2-1-22）式作如下变换，令

$$\sigma_i=\sqrt{s_{ii}} \qquad (i=1,\ 2,\ \cdots,\ m,\ y)$$

$$b_i=\frac{\sigma_y}{\sigma_i}b_i' \qquad (i=1, 2, \cdots, m)$$

$$b_0=b_0'$$

$$r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}=\frac{s_{ij}}{\sigma_i\sigma_j} \quad (i, j=1, 2, \cdots, m, y)$$

于是，（2-1-22）式可以变换为

$$\begin{cases} r_{11}b_1'+r_{12}b_2'+\cdots+r_{1m}b_m'=r_{1y} \\ r_{21}b_1'+r_{22}b_2'+\cdots+r_{2m}b_m'=r_{2y} \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ r_{m1}b_1'+r_{m2}b_2'+\cdots+r_{mm}b_m'=r_{my} \end{cases} \qquad (2\text{-}1\text{-}29)$$

和 $$b_0'=\bar{y}-\frac{\sigma_y}{\sigma_1}b_1'\bar{x}_1-\frac{\sigma_y}{\sigma_2}b_2'\bar{x}_2-\cdots-\frac{\sigma_y}{\sigma_m}b_m'\bar{x}_m \qquad (2\text{-}1\text{-}30)$$

由（2-1-29）式可构成一个增广矩阵

$$R=[r_{ij}]_{m\times(m+1)} \qquad (i, j=1, 2, \cdots, m, y)$$

逐步回归计算过程中，引进一个变量或者是剔除一个变量，都要对增广矩阵R作一次变换。

由观测数据算得的矩阵$R^{(0)}=[r_{ij}^{(0)}]$称为原始增广矩阵。这里约定$R^{(t)}=[r_{ij}^{(t)}]$表示第t次变换后的矩阵。显然，$R^{(t)}$是由$R^{(t-1)}$变换而来。

矩阵变换时，各元素的计算公式如下

$$r_{ij}^{(t)}=\begin{cases} r_{kj}^{(t-1)}/r_{kk}^{(t-1)} & (\text{在}k\text{行上的元素，}i=k,\ j\neq k) \\ r_{ij}^{(t-1)}-r_{ik}^{(t-1)}r_{kj}^{(t-1)}/r_{kk}^{(t-1)} & (\text{在其他行列上的元素，}i\neq k, j\neq k) \\ -r_{ik}^{(t-1)}/r_{kk}^{(t-1)} & (\text{在}k\text{列上的元素，}i\neq k,\ j=k) \\ 1/r_{kk}^{(t-1)} & (\text{在}k\text{行}k\text{列交叉点的元素，}i=k, j=k) \end{cases} \qquad (2\text{-}1\text{-}31)$$

（2-1-31）式来源于求解线性方程组的矩阵变换公式，有如下性质

（1）$$r_{ij}^{(t)}=\begin{cases} r_{ji}^{(t)} & (i,\ j\text{都是或都不是}t\text{步变量的下标}) \\ -r_{ji}^{(t)} & (i,\ j\text{中有一个是}t\text{步变量的下标}) \end{cases}$$

所以，在计算矩阵$R^{(t)}$时，只要算出主对角线上以及右上三角部分的各元素就可以了。

（2）$R^{(t)}$只与第t步已引进方程的全体变量有关，而与这些变量引进的次序无关，也与已剔除的变量无关。

（3）对同一变量x_i施行两次变换，矩阵复原，即$R^{(t)}=R^{(t-2)}$。因此，第t步无论是引进或剔除变量x_i，都是对前一步矩阵$R^{(t-1)}$施行关于变量x_i的矩阵变换。如果x_i已在回归方程中，则t步变换就是从方程中剔除变量x_i；如果x_i不在回归方程中，则矩阵变换的结果是将x_i引进到方程中。

二、逐步回归的计算步骤

1. 准备工作

（1）由原始数据构成增广矩阵

$$X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \\ x_{m+1\ 1} & x_{m+1\ 2} & \cdots & x_{m+1\ n} \end{pmatrix}$$

矩阵中的$x_{m+1\ 1}=y_1$，$x_{m+1\ 2}=y_2$，…，$x_{m+1\ n}=y_n$。

（2）选定检验临界值$F_{进}$与$F_{出}$

当样品的数量n较大时，由于引进变量时的$(n-p-2)$与剔除变量时的（$n-p-1$）相差很小，所以可取$F_{进}=F_{出}=F^*$。F^*的值要根据实际情况确定，若欲使回归方程中引进较多的变量，则F^*的值不宜过高，或者检验水平α的值不宜过小。

（3）由原始数据的增广矩阵计算相关系数矩阵$R^{(0)}=[r_{ij}]_{m\times(m+1)}$，其中

$$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}} \quad (i,\ j=1,\ 2,\ \cdots,\ m,\ m+1)$$

$$S_{ij}=\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j) \qquad (i,\ j=1,\ 2,\ \cdots,\ m,\ m+1)$$

从而可以得到初始相关系数矩阵$R^{(0)}$。

2. 逐步回归的前三步计算

（1）选择第一个变量进入方程，由$R^{(0)}$计算

$$W_i^{(0)}=\frac{(r_{iy}^{(0)})^2}{r_{ii}^{(0)}} \qquad (i\leqslant m)$$

从中找出最大的$W_i^{(0)}$，即

$$W_{k1}^{(0)}=\max_{1\leqslant i\leqslant m} W_i^{(0)}$$

再计算

$$F_{1k1}=\frac{W_{k1}^{(0)}(n-2)}{y_{yy}-W_{k1}^{(0)}}$$

若 $F_{1k1}>F^*$，则把地质变量x_{k1}引进回归方程中。并按（2-1-31）式对相关系数矩阵进行变换，从而得到$R^{(1)}=[r_{ij}^{(1)}]$。若$F_{1k}\leqslant F^*$，则停止计算，即在F^*检验水平下所有的变量都不能进入回归方程中。

（2）选择第二个变量进入回归方程，是由$R^{(1)}$计算

$$W_i^{(1)}=\frac{(r_{iy}^{(1)})^2}{r_{ii}^{(1)}} \quad (i\leqslant m,\ i\neq k1)$$

从中找出最大的$W_i^{(1)}$，即

$$W_{k2}^{(1)}=\max_{\substack{1\leqslant i\leqslant m\\ i\neq k1}} W_i^{(1)}$$

若
$$F_{1k2}^{(1)}=\frac{W_{k2}^{(1)}(n-2-1)}{r_{yy}^{(1)}-W_{k2}^{(1)}}>F^*$$

则把变量x_{k2}引进回归方程中，并按（2-1-31）式对相关系数矩阵进行变换，而得到 $R^{(2)}=[r_{ij}^{(2)}]$。

（3）前面两步已引进两个变量进入了回归方程。所以，下一步将要考虑在已引进的两个变量中，是否有要剔除的变量。如果不能剔除，则可以再引进新的变量。这里有三种情况

①如果 $F_{2k1}^{(2)}=\dfrac{W_{k1}^{(2)}(n-2-1)}{r_{yy}^{(2)}}\leqslant F^*$

则要从回归方程中剔除变量x_{k1}，并且计算$R^{(3)}=[r_{ij}^{(3)}]$。

②如果 $F_{2k1}^{(2)}=\dfrac{W_{k1}^{(2)}(n-2-1)}{r_{yy}^{(2)}}>F^*$

则不能剔除x_{k1}，并计算

$$W_i^{(2)}=\frac{(r_{iy}^{(2)})^2}{r_{ii}^{(2)}}\quad (i\leqslant m;\ i\neq k1,\ k2)$$

找出最大的$W_i^{(2)}$，即

$$W_{k3}^{(2)}=\max_{\substack{1<i<m\\ i\neq k1,k2}} W_i^{(2)}$$

若

$$F_{1k3}^{(3)}=\frac{W_{k3}^{(3)}(n-3-1)}{r_{yy}^{(3)}-W_{k3}^{(3)}}>F^*$$

则把变量x_{k3}引进到方程中，并计算$R^{(3)}=[r_{ij}^{(3)}]$。

③如果已经再没有可以进入方程的新变量，即$F_{1k3}^{(3)}\leqslant F^*$时，则停止计算。就是说在$F^*$检验水平下只能使两个变量进入回归方程。最后计算

$$b_i=b_i^{(2)}=\frac{\sigma_y}{\sigma_{ki}}r_{kiy}^{(2)}\quad (i=1,\ 2)\qquad (2\text{-}1\text{-}32)$$

$$b_0=\bar{y}-\sum_{i=1}^{2}b_i\bar{x}_{ki}\qquad (i=1,\ 2)\qquad (2\text{-}1\text{-}33)$$

将b_0，b_i代入回归方程，则有

$$\hat{y}=b_0+b_1x_{k1}+b_2x_{k2}$$

复相关系数为 $R=\sqrt{1-r_{yy}^{(2)}}$

剩余标准差为

$$s_y=\sigma_y\sqrt{\frac{r_{yy}^{(2)}}{n-2-1}}$$

3. 第三步以后的逐步回归计算

对于上述的①、②情况下，需要由$R^{(3)}$继续进行逐步回归计算。那么，在一般情况下，如果逐步回归计算到t步，则由$R^{(t)}$进行（$t+1$）步计算时，也有三种情况。

（1）对所有的t步变量x_i都要计算

$$W_i^{(t)}=\frac{(r_{iy}^{(t)})^2}{r_{ii}^{(t)}}$$

从中找出最小的$W_i^{(t)}$，即

$$W_k^{(t)}=\min_{1\leqslant i\leqslant p}W_i^{(t)}$$

若

$$F_{2k}^{(t)}=\frac{W_k^{(t)}(n-p-1)}{r_{kk}^{(t)}}\leqslant F^*$$

则在第（$t+1$）步把变量x_k剔除，并计算（$t+1$）步的$R^{(t+1)}=[r_{ij}^{(t+1)}]$。

（2）若$F_{2k}^{(t)}>F^*$，则对所有还没有进入回归方程中的变量计算

$$W_i^{(t)}=\frac{(r_{iy}^{(t)})^2}{r_{ii}^{(t)}}\quad(i\leqslant m;\ i\neq k1,k2,\cdots,kp)$$

从中找出最大的$W_i^{(t)}$，即

$$W_s^{(t)}=\max_{\substack{1\leqslant i\leqslant m\\ i\neq k1,k2,\cdots kp}}W_i^{(t)}$$

若

$$F_{1s}^{(t)}=\frac{W_s^{(t)}(n-p-2)}{r_{yy}^{(t)}-W_s^{(t)}}>F^*$$

则把变量x_s引进到回归方程之中，并计算$R^{(t+1)}=[r_{ij}^{(t+1)}]$

（3）如果所有的还没有进入回归方程中的变量，在$i\neq k1,k2,\cdots,kp$时变量的$F_{1i}^{(t)}\leqslant F^*$时，则逐步回归计算结束。也就是说在$F^*$检验水平下只能有$p$个变量$x_{k1},x_{k2},\cdots,x_{kp}$进入回归方程之中。

最后，计算出回归方程的待定系数

$$b_i=b_i^{(t)}=\frac{\sigma_y}{\sigma_{ki}}r_{ki,y}\quad(i=1,\ 2,\ \cdots,\ p)\tag{2-1-34}$$

$$b_0=\bar{y}-\sum_{i=1}^{p}b_i\bar{x}_{ki}\tag{2-1-35}$$

将b_0，b_i代入回归方程，则有

$$\hat{y}=b_0+b_1x_{k1}+b_2x_{k2}+\cdots+b_px_{kp}$$

复相关系数为　$R=\sqrt{1-r_{yy}^{(t)}}$

剩余标准差为　$s_y=\sigma_y\sqrt{\frac{r_{yy}^{(t)}}{n-p-1}}$

三、算　　例

［1］康南（Connan）用表2-1-4中12个地区的生油层数据，计算出大量生油门限值的回归方程式为

$$\ln t=6942\left(\frac{1}{T+273}\right)-14.965\tag{2-1-36}$$

除表2-1-4中的数据外，再增加6组我国生油层数据，见表2-1-5。取实际深度（H）、现在温度（T）、$\frac{1}{H}$、$\frac{1}{T+273}$、H^2、T^2共6个变量，以地层年龄t作为因变量，进行多元逐步回归计算，结果选入$\frac{1}{T+273}$及$\frac{1}{H}$两个变量，得出的回归方程为

$$\ln t = 7070\left(\frac{1}{T+273}\right) - 2170\left(\frac{1}{H}\right) - 14.64 \qquad (2\text{-}1\text{-}37)$$

表2-1-4　十二个盆地（地区）的生油层数据

序号	盆地（地区）	门限值		
		地层年龄（t）（10^6a）	现在温度（T）（°C）	实际深度（H）（m）
1	杜阿拉盆地(喀麦隆)	70	65	1200
2	洛山矶盆地（美国）	12	115	2440
3	文吐拉盆地（美国）	12	127	2740
4	巴黎盆地（法国）	180	60	1400
5	阿启坦盆地（1）（法国）	112	90	3300
6	阿启坦盆地（2）（法国）	135	72	2600
7	卡马尔圭盆地（法国）	38	106	3250
8	阿尤恩地区	105	85	2740
9	苏禄海盆地（沙巴）	12	120	3050
10	塔拉纳基盆地（新西兰海上）	70	80	2900
11	亚马逊盆地（委内瑞拉）	359	62	1750
12	塔拉纳基盆地（新西兰海上）	32	95	3350

表2-1-5　我国的六组生油层数据

序号	盆地(地区)	门限值		
		地层年龄(t)（10^6a）	现在温度(T)（°C）	实际深度(H)(m)
13	东营盆地	35	93	2200
14	潜江盆地	35	90	2200
15	松辽盆地(1)	110	70	1380
16	松辽盆地(2)	100	65	1230
17	松辽盆地(3)	90	63	1180
18	辽河盆地	50	81	1700

复相关系数$R=0.931$。

把表2-1-4、表2-1-5中的原始数据T及H分别代入方程式（2-1-36）与（2-1-37）式中，计算得到最大生油门限时间和误差如表2-1-6所示。按

$$误差 = \left|\frac{计算的生油门限时间 - 实际地层年龄}{实际地层年龄}\right| \times 100\%$$

计算结果表明，表2-1-6中前12个地区的数据，按康南得出（2-1-36）式的平均误差为

58.7%，而（2-1-37）式的平均误差为29.7%。表2-1-6中后6个我国的生油层数据，按康南得出（2-1-36）式的平均误差为119.8%；而(2-1-37)式的平均误差为18.3%显然,(2-1-37)式的误差大大低于康南得出的方程式（2-1-36）式。这就是说，深度因素是不可忽略的。

从（2-1-37）式还可以看出，生油层温度高、埋藏浅将使生油门限时间短；反之，生油层温度低、埋藏深则生油门限时间长。这说明地温因素也是重要的。

表2-1-6 两个方程的误差比较

序号	地层年龄 (10^6a)	康南方程的预测值 (10^6a)	误差 (%)	(2-1-37)式的预测值 (10^6a)	误差 (%)
1	70	263.34	276.9	8[illegible].25	24.6
2	12	18.67	55.6	14.76	23
3	12	10.91	9.1	9.48	21
4	180	358.47	99.2	156.02	13.3
5	112	64.00	42.9	65.36	41.6
6	135	173.59	28.6	146.16	8.[illegible]
7	38	28.55	24.9	28.56	25
8	105	83.60	20.4	75.32	28.[illegible]
9	12	14.87	23.9	13.99	16.6
10	70	110.02	57.2	[illegible]03.6[illegible]	47.9
11	359	316.52	11.8	186.79	48
12	32	49.35	54.2	5[illegible]	69.1
13	35	54.72	56.3	40.04	14.4
14	35	64	82.8	46.99	34.3
15	110	195.20	77.4	76.76	30.2
16	100	263.34	163.3	91.16	8.8
17	90	297.58	230.6	95.84	6.5
18	50	104.08	108.2	[illegible].75	[illegible]5.4

［2］ 有人曾对美国加利福尼亚州作过每一mile²❸面积内的油气产值估计，所用的方法是多元逐步回归分析方法。回归方程中最终引进了10个变量，其中2个为经济变量，8个为地质变量。这些变量与每mile²面积内油气产值Y的回归方程如下：

$$
\begin{aligned}
Y= & -3.046+0.0007925x_{11}+0.9740x_{12}+0.05233x_{13} \\
& +0.02688x_{15}+0.04462x_{16}+0.6388x_{17}+0.1460x_{18} \\
& +0.005240x_{19}+0.003471x_{21}-0.03671x_{23}
\end{aligned}
$$

式中 Y——每mile²面积内油气产值（1968年，美元）的对数值（$\log_{10}$）；

x_{11}——美国的国家总产值（1968年，美元）；

x_{12}——每mile²面积内人口的对数值（$\log_{10}$）；

x_{13}——新生代海相沉积岩的百分比；

x_{15}——新生代非海相沉积岩的百分比；

x_{16}——中生代海相沉积岩的百分比；

❸ mile² = 2.589988×10⁶m²。

x_{17}——古生代海相沉积岩的百分比；

x_{18}——前白垩系火山岩及火山变质岩百分比；

x_{19}——中生代花岗岩百分比乘中生代海相岩石百分比；

x_{21}——中生代花岗岩百分比乘前白垩系火山岩及变质岩百分比；

x_{23}——重力值（平均布格异常）。

根据回归方程可以绘出油气产值的等值线及趋势残差图。研究者认为对低偏差值地区要给以足够的重视，因为对这些地区研究得最差，开发程度也最低。

第二章 趋势分析

许多地质现象，例如某一地层界面的空间展布，某种地球化学元素的区域性变化，某个储集层中流体的压力变化等等都可以用一个空间曲面描述。趋势分析的出发点，就是要把所研究的地质现象进行分解，即

$$z=\hat{z}+A+R \qquad (2\text{-}2\text{-}1)$$

式中 z——地质现象（空间曲面）；

$\hat{z}$——区域性因素（背景值）；

A——局部性因素（异常值）；

R——随机性因素（干扰值）。

地质研究工作中，特别是与找矿有关的地质问题，往往都可归结为寻找区域上的异常值问题。趋势分析的出发点就是以某一人为构造的数学曲面，即趋势面代替区域性因素（背景值），同时再用统计方法消除随机性因素（干扰值），以达到突出局部性因素（异常值）之目的。

例如，某个勘探地区的油气地质演化过程中，曾有一套生油岩系沉降到生油门限深度以下，即进入了油气生成期。此时，如果该探区发生了构造变动，并且形成了各种类型的地质圈闭，这些地质圈闭是油气生成期的同生圈闭，最容易捕集油气。因而成为重点勘探对象，见图2-2-1的①。

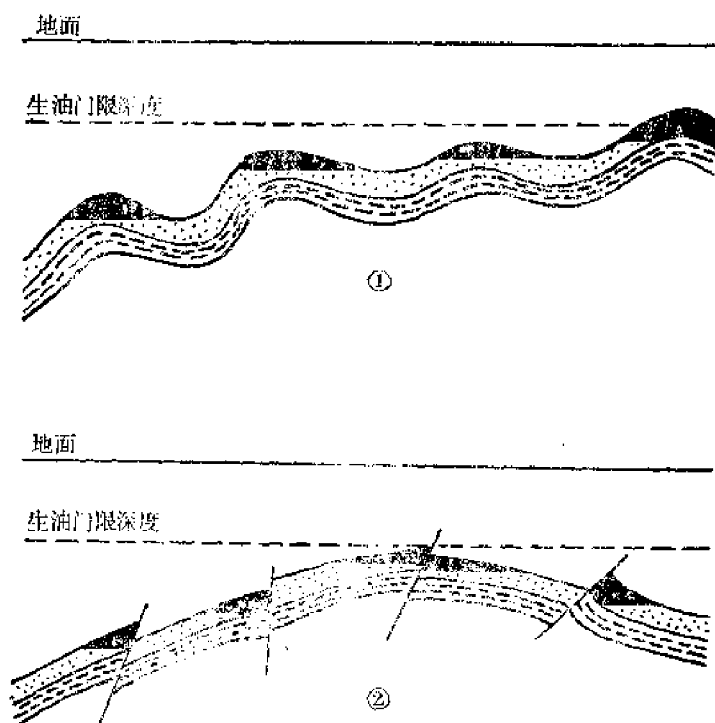


图2-2-1 生油期同生圈闭

但是，由于经过后来的多期构造变动，使原来圈闭形态发生了程度不同的变化。所以，目前通过各种勘探手段获得的实测圈闭图，根本反映不出或者不能完全反映出生油时期同生圈闭的形态特征，见图2-2-1的②。

在这种情况下，有可能通过趋势分析以某一数学曲面，即趋势面代替油气藏形成后的构造演变趋势，以达到恢复或者部分恢复生油同期的圈闭形态之目的。这显然对寻找有利勘探地带是有益处的。

目前，人为构造趋势面的数学方法有两种，一种是多项式函数，另一种是傅立叶级数。

第一节　多项式趋势分析

多项式构造趋势面是目前最常用的方法，一次多项式表示的趋势面是空间的一个平面；二次趋势面是抛物面、椭球面或双曲面；三次及三次以上的趋势面是形态复杂的空间曲面；随着趋势面的次数增高，曲面的形态就越复杂。

一、多项式趋势面方程的建立

如果有一组总共n个观测数据，其观测点的平面坐标为（x_i，y_i），地质变量的观测值为z_i。对于这组观测数据的一次趋势面方程为

$$\hat{z}_i=b_0+b_1x_i+b_2y_i \quad (i=1,2,\cdots n) \qquad (2\text{-}2\text{-}2)$$

式中　$\hat{z}_i$——第i个观测点的趋势值；

b_0，b_1，b_2——一次趋势面方程的待定系数。

二次趋势面方程为

$$\hat{z}_i=b_0+b_1x_i+b_2y_i+b_3x_i^2+b_4x_iy_i+b_5y_i^2 \qquad (2\text{-}2\text{-}3)$$

$$(i=1,\ 2,\ \cdots n)$$

二次趋势面方程中的待定系数有b_0，b_1，b_2，b_3，b_4，b_5共6个。趋势面方程的次数越高，待定系越多。

为使趋势面最大限度地逼近原始观测数据，可采用最小二乘法使每个观测点的观测值与趋势值之差(残差)的平方和最小，即

$$\min G=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2$$

对于一次趋势面则为

$$\min G=\sum_{i=1}^{n}(z_i-b_0-b_1x_i-b_2y_i)^2$$

为使G最小，可按求极值方法使b_0，b_1，b_2对G的偏导数为0，亦即

$$\begin{cases}\dfrac{\partial G}{\partial b_0}=2\sum\limits_{i=1}^{n}(z_i-b_0-b_1x_i-b_2y_i)(-1)=0\\[2ex]\dfrac{\partial G}{\partial b_1}=2\sum\limits_{i=1}^{n}(z_i-b_0-b_1x_i-b_2y_i)(-x_i)=0\\[2ex]\dfrac{\partial G}{\partial b_2}=2\sum\limits_{i=1}^{n}(z_i-b_0-b_1x_i-b_2y_i)(-y_i)=0\end{cases}$$

经整理，可以得到如下正规方程组

$$\begin{cases} b_0\Sigma 1+b_1\Sigma x+b_2\Sigma y=\Sigma z \\ b_0\Sigma x+b_2\Sigma x^2+b_2\Sigma xy=\Sigma zx \\ b_0\Sigma y+b_2\Sigma xy+b_2\Sigma y^2=\Sigma zy \end{cases}$$

写成矩阵形式则为

$$\begin{pmatrix} \Sigma 1 & \Sigma x & \Sigma y \\ \Sigma x & \Sigma x^2 & \Sigma xy \\ \Sigma y & \Sigma xy & \Sigma y^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \Sigma z \\ \Sigma zx \\ \Sigma zy \end{pmatrix}$$

此系数矩阵为正定、对称矩阵，可用高斯消元法求解联立方程组或用系数矩阵求逆方法求出待定系数b_0，b_1，b_2。将系数回代到（2-2-2）式后，则可得到趋势方程，并且可将观测点的平面位置坐标代入方程，求出所有观测点的趋势值$\hat{z}$，从而可以绘制趋势面的平面等值线图。当然，用趋势面方程可以计算出研究地区中任意一点p的趋势值$\hat{z}_p$。

二次趋势面方程的待定系数有6个，即b_0，b_1，b_2，b_3,b_4，b_6，求待定系数的方法与一次趋势面完全类似，即可使每个待定系数对残差平方和的偏导数为0，经过整理，可以得到如下形式的正规方程组，写成矩阵形式则为

$$\begin{pmatrix} \Sigma 1 & \Sigma x & \Sigma y & \Sigma x^2 & \Sigma xy & \Sigma y^4 \\ \Sigma x & \Sigma x^2 & \Sigma xy & \Sigma x^3 & \Sigma x^2y & \Sigma xy^2 \\ \Sigma y & \Sigma xy & \Sigma y^2 & \Sigma x^2y & \Sigma xy^2 & \Sigma y^3 \\ \Sigma x^2 & \Sigma x^3 & \Sigma x^2y & \Sigma x^4 & \Sigma x^3y & \Sigma x^2y^2 \\ \Sigma xy & \Sigma x^2y & \Sigma xy^2 & \Sigma x^3y & \Sigma x^2y^2 & \Sigma xy^3 \\ \Sigma y^2 & \Sigma xy^2 & \Sigma y^3 & \Sigma x^2y^2 & \Sigma xy^3 & \Sigma y^4 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} \Sigma z \\ \Sigma zx \\ \Sigma zy \\ \Sigma zx^2 \\ \Sigma zxy \\ \Sigma zy^2 \end{pmatrix}$$

求解方程组则可算出待定系数b_0，b_1，b_2，b_3，b_4，b_6。回代到（2-2-3）式便可以得到二次趋势面方程。

趋势分析的次数越高，系数矩阵的元素就越多，其中有些元素的方次也越来越大。趋势分析的实质是一种多维（二维或三维）的高次非线性回归分析，二次趋势面包括了一次趋势面成分；（$n+1$）次的趋势面包括了所有低于（$n+1$）次的趋势面成分。

二维趋势面是研究变量在平面上的变化趋势，而三维趋势面是研究变量在空间上的变化趋势，所以除考虑横坐标x，纵坐标y以外，还要考虑一个高程（深度）坐标h，计算趋势面待定系数的方法与二维趋势面类似，只是趋势方程的项数更多，计算时更复杂。

二、趋势面的拟合度

趋势面的次数越高，对原始数据观测值的逼近程度也越好。在数学上可用离差平方和表示趋势面对观测值的拟合程度，即

$$S=\sum_{i=1}^{n}(z_i-\bar{z})^2 \qquad (2\text{-}2\text{-}4)$$

式中 z_i——第i个观测点的变量观测值；

$\bar{z}$ ——n个观测点的变量平均值。

上式可以分解为两部分，即

$$S=\sum_{i=1}^{n}(z_i-\bar{z})^2=\sum_{i=1}^{n}[(z_i-\hat{z}_i)+(\hat{z}_i-\bar{z})]^2$$

$$=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2+2\sum_{i=1}^{n}(z_i-\hat{z}_i)(\hat{z}_i-\bar{z})+\sum_{i=1}^{n}(\hat{z}_i-\bar{z})^2$$

其中交叉项

$$2\sum_{i=1}^{n}(z_i-\hat{z}_i)(\hat{z}_i-\bar{z})=0$$

所以

$$S=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2+\sum_{i=1}^{n}(\hat{z}_i-\bar{z})^2$$

令

$$V=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2 \quad ; \quad U=\sum_{i=1}^{n}(\hat{z}_i-\bar{z})^2$$

则有 $S=V+U$

式中的V为观测值与趋势值之差，称为残差平方和，V越大，拟合程度越低。U为趋势值与观测值之差，称为回归平方和，U越大，拟合程度越高。令

$$C=\frac{U}{S}=\left(1-\frac{V}{S}\right)=\left(1-\frac{\sum_{i=1}^{n}(z_i-\hat{z}_i)^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2}\right)\times 100\% \qquad (2\text{-}2\text{-}5)$$

C可作为衡量拟合度的标准。$C=100\%$时，表示所有趋势值与观测值完全一致，即趋势面通过所有的实际观测值；$C=0\%$时，表示趋势面与观测值完全不拟合，这种情况一般不容易出现；$C=70\%$时，表示趋势面反映了实际观测值的70%，还有30%在趋势面中没有反映出来。

对于所建立的趋势面是否有意义，从数学上可用下面的统计量进行辅助性检验。

$$F=\frac{\dfrac{U}{m}}{\dfrac{V}{n-m-1}} \qquad (2\text{-}2\text{-}6)$$

统计量F服从$F_\alpha(m,\ n-m-1)$分布，

(2-2-6)式中 n——样品个数；

m——不包括b_0在内的多项式趋势面方程的系数个数。

当给定检验水平α后，可由F分布表查出临界值F_α，当$F>F_\alpha$时，则认为趋势面所反映变量的变化情况是显著的；当$F\leqslant F_\alpha$时，则认为不显著。

三、残差分析

前已述及，趋势分析是假设我们所研究的地质现象可以分解为三个部分，即趋势值、异常值和干扰值。

各点的观测值减去趋势值为残差值Δz_i，也可称为剩余值，即

$$\Delta z_i = z - \hat{z}_i = A_i + R_i$$

这就是说，残差值是由异常值A和干扰值R两部分组成。对于矿产预测来说，一般情况下是正残差才有意义，例如，对于油气勘探来说，正向构造圈闭才有意义，又如对于分散性的金属找矿来说品位为正异常才有意义。在此，正残差可表示为Δz_i^+，即

$$\Delta z_i^+ = A_i + R_i$$

对于绝大多数的随机干扰值来说，其分布概型为正态分布。因而，为了消除随机性的干扰值，经常采用如下的简单处理方法，即以n个观测值中的m个正残差值的平均值

$$\overline{\Delta z^+} = \frac{1}{m} \sum_{i=1}^{n} \Delta z_i^+$$

来代替每个观测点上的R_i。那么，从残差值中减掉$\overline{\Delta z^+}$之后就得到了正的异常值A_i，即

$$A_i = \Delta z_i^+ - \overline{\Delta z^+}$$

对于许多地质问题，研究人员所关心的只是趋势值之间的相对关系，例如，对于找矿来说异常值高区总比低区为好。因而也经常令$A_i = \Delta z_i$。

关于随机干扰值的处理，还有其他许多方法，这里就不一一叙述了。

通过趋势分析得到异常值A_i之后，便可以绘制异常值的等值线图，从图上可以确定出异常区带，以指导找矿勘探工作。

四、算　例

某探区钻探了52口探井，井位分布见图2-2-2。

表2-2-1中列出了52口探井的横坐标x、纵坐标y以及某含油层顶面的海拔高程z。表中

表2-2-1　钻井井位坐标及油层顶面海拔高程数据表

序号	横坐标，x	纵坐标，y	海拔高程，z(m)	序号	横坐标，x	纵坐标，y	海拔高程，z(m)
1	0.3	6.4	870	19	4.1	4.6	760
2	1.4	6.7	783	20	4.9	4.2	790
3	2.4	6.1	755	21	6.3	4.3	820
4	3.6	6.2	690	22	0.9	3.7	855
5	5.7	6.2	800	23	1.7	3.8	812
6	1.6	5.7	800	24	2.4	3.8	778
7	2.9	5.1	730	25	3.7	3.5	817
8	3.4	5.3	728	26	4.5	3.2	827
9	3.4	5.7	710	27	5.2	3.2	805
10	4.8	5.6	780	28	6.3	3.4	840
11	5.3	5.0	804	29	0.3	2.4	890
12	6.2	5.2	855	30	2.0	2.7	830
13	0.2	4.7	830	31	3.8	2.3	873
14	0.9	4.2	813	32	6.3	2.2	875
15	2.3	4.8	762	33	0.6	1.7	873
16	2.5	4.5	765	34	1.5	1.8	865
17	3.0	4.5	740	35	2.1	1.5	841
18	3.5	4.5	765	36	2.1	1.1	832

续表

序号	横坐标，x	纵坐标，y	海拔高程，z(m)	序号	横坐标，x	纵坐标，y	海拔高程，z(m)
37	3.1	1.1	908	45	2.1	0.7	880
38	4.5	1.8	855	46	2.3	0.3	870
39	5.5	1.7	850	47	3.1	0.01	880
40	5.7	1.6	882	48	4.1	0.8	960
41	6.2	1.0	910	49	5.4	0.4	890
42	0.4	0.5	970	50	6.0	0.1	860
43	1.4	0.6	915	51	5.7	3.0	830
44	1.4	0.1	890	52	3.6	6.6	705

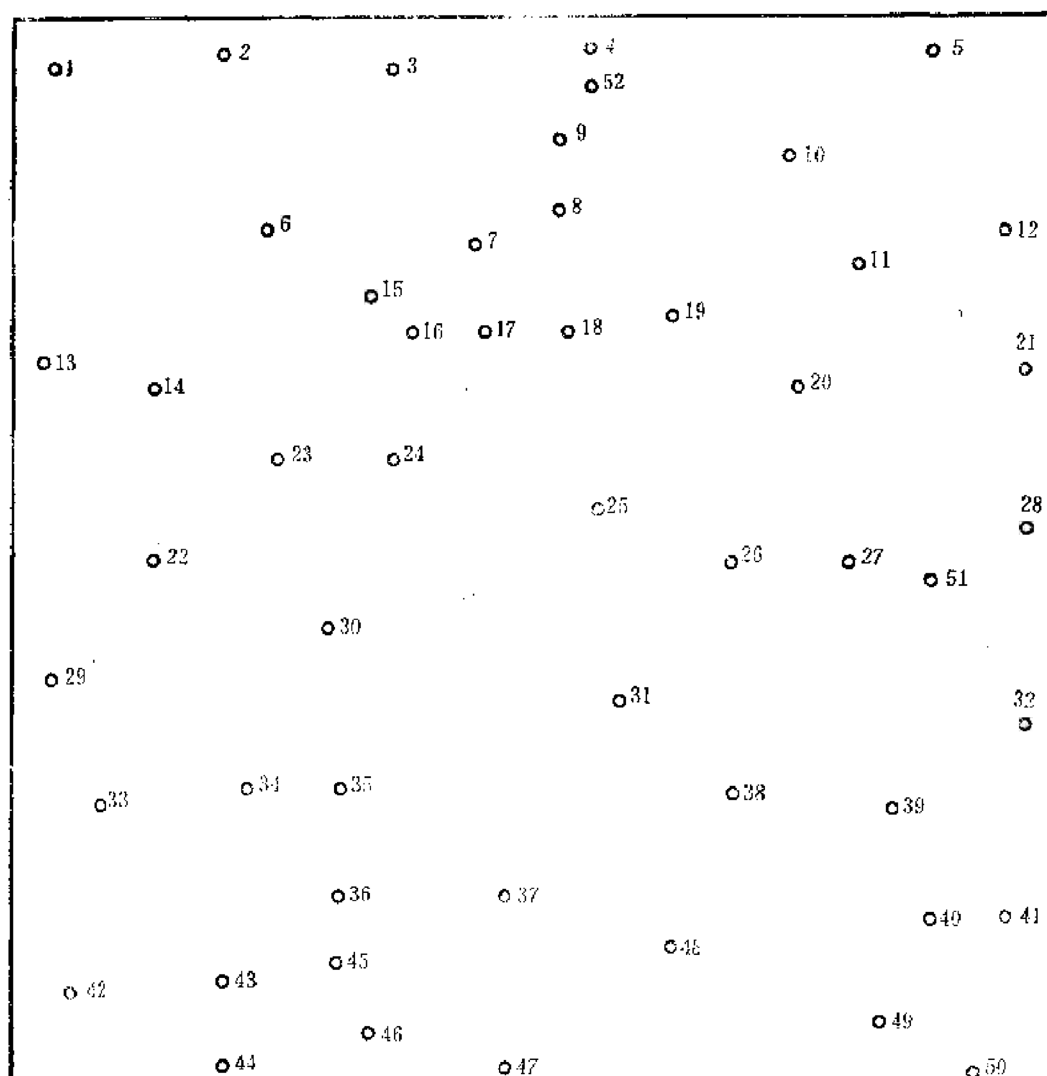


图2-2-2　钻井井位图

的井位坐标是以图2-2-2的左下角为坐标原点（0，0）实际测量的。含油层海拔高度的单位是m。按平面插值法生成的构造等值线图见图2-2-3。

经计算1次至5次的趋势面图及残差图见图2-2-4至图2-2-13。从这些图中可以发现，随着趋势面方程次数的增加，拟合程度越来越高。从各次的残差图上可以发现一些正向局部构造；且随着拟合度增高，局部构造形态也在发生变化。如果结合实际地质资料，仔细研究这些残差图对于指导油气勘探是有益处的。

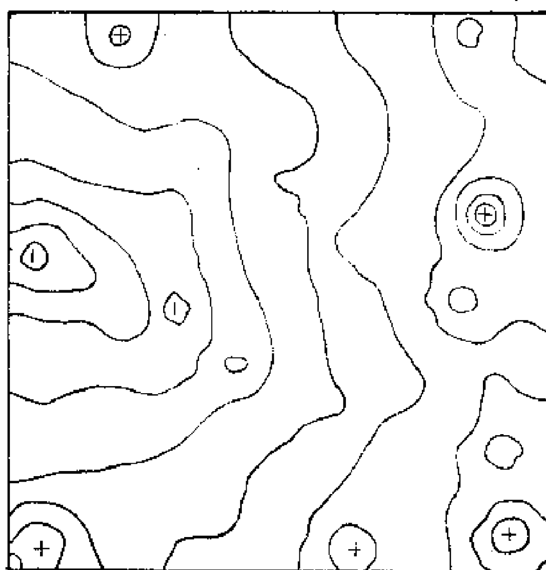


图2-2-3 含油层顶面平面插值图

一次趋势面方程为

$$\hat{z}=913.825-1.696x-25.257y$$

拟合度 $C=0.657$

二次趋势面方程为

$$\hat{z}=976.376-52.378x-30.437y+7.334x^2+0.354xy+0.873y^2$$

拟合度 $C=0.796$

三次趋势面方程为

$$\hat{z}=908.434-13.676x+16.682y+6.268x^2-14.837xy-11.862y^2-0.836x^3+2.675x^2y-0.427xy^2+1.483y^3$$

拟合度 $C=0.890$

四次趋势面方程为

$$\hat{z}=977.753-121.906x+19.990y+43.287x^2+38.475xy-47.295y^2-4.994x^3-10.288x^2y-5.985xy^2+12.139y^3+0.103x^4+0.909x^3y+0.685x^2y^2+0.100xy^3-0.864y^4$$

拟合度 $C=0.924$

五次趋势面方程为

$$\hat{z}=910.069+9.057x+31.270y-26.768x^2-65.719xy+6.854y^2+16.030x^3+19.394x^2y+15.341xy^2-14.402y^3-3.634x^4-0.134x^3y-6.949x^2y^2+0.659xy^3+3.301y^4+0.271x^5-0.174x^4y+0.498x^3y^2+0.293x^2y^3-0.199xy^4-0.213y^5$$

拟合度 $C=0.924$

第二节 调和趋势分析

许多地质现象往往具有明显的周期性变化特征，例如，某个地层界面的波状起伏，某项地球化学指标在区域上的周期性变化，古地磁的周期性变化，沉积旋回，构造旋回，岩浆活动旋回等等都具有明显的周期性变化特征。调和趋势分析的出发点是用简单的波形叠加来表示地质问题中的复杂波形，也就是用傅立叶级数构造的趋势面去拟合具有波状变化的地质观测数据，以适应所研究地质问题的波状变化特征。

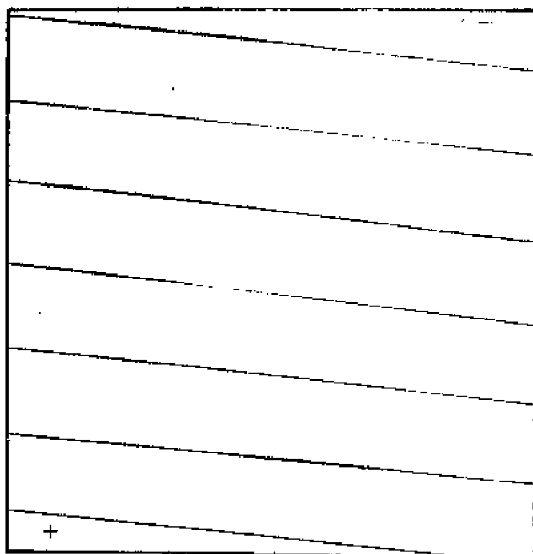


图2-2-4　一次趋势面图

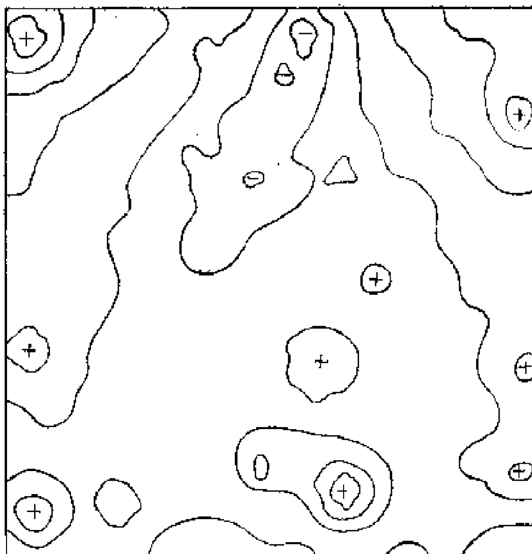


图2-2-5　一次残差图

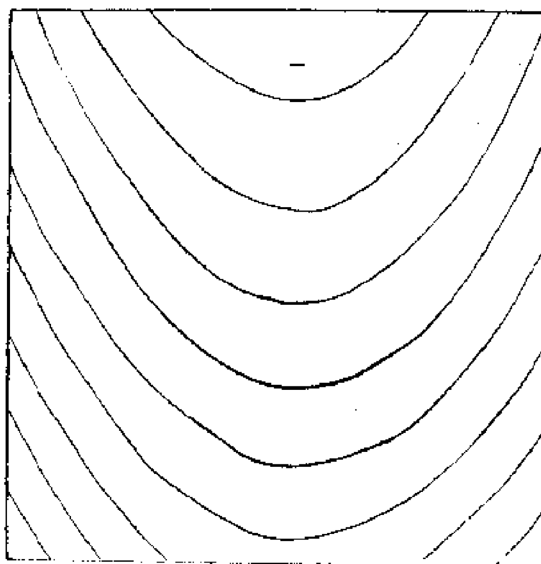


图2-2-6　二次趋势面图

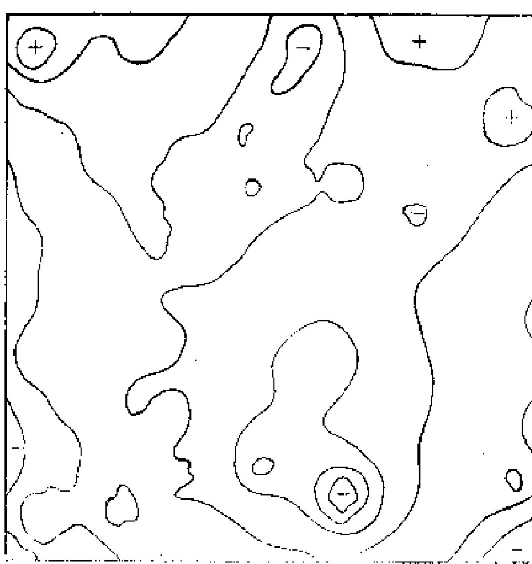


图2-2-7　二次残差图

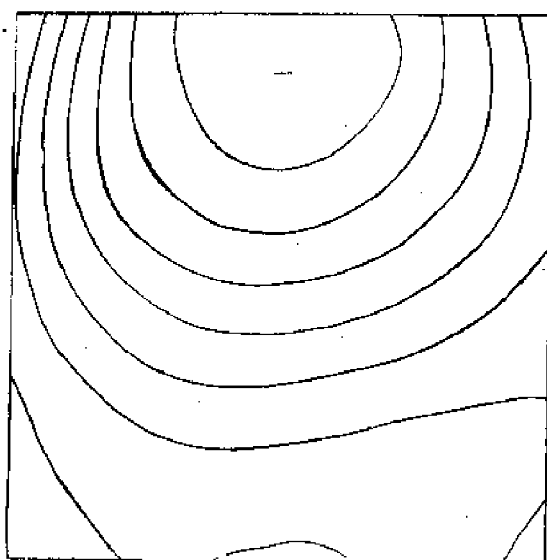


图2-2-8　三次趋势面图

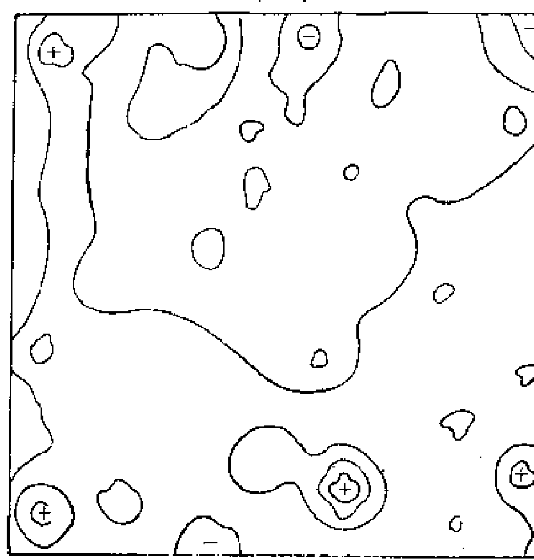


图2-2-9　三次残差图

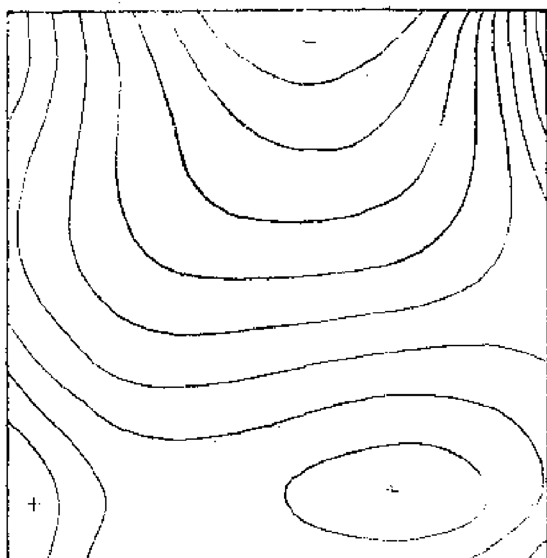


图2-2-10　四次趋势面图

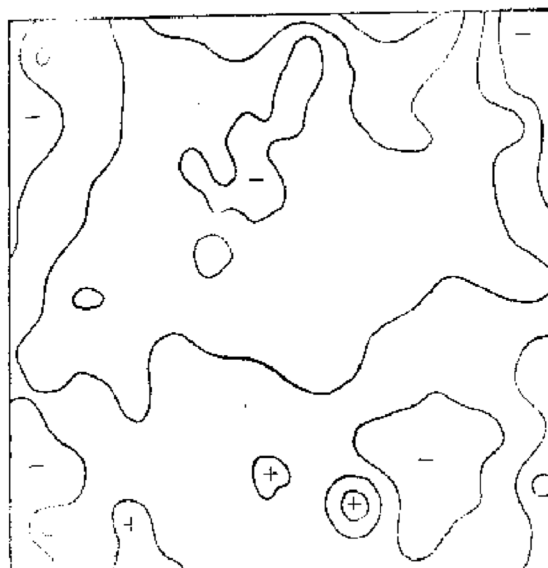


图2-2-11　四次残差图

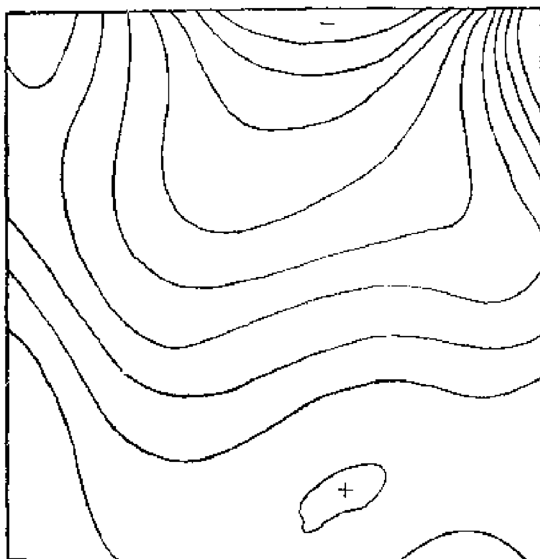


图2-2-12　五次趋势面图

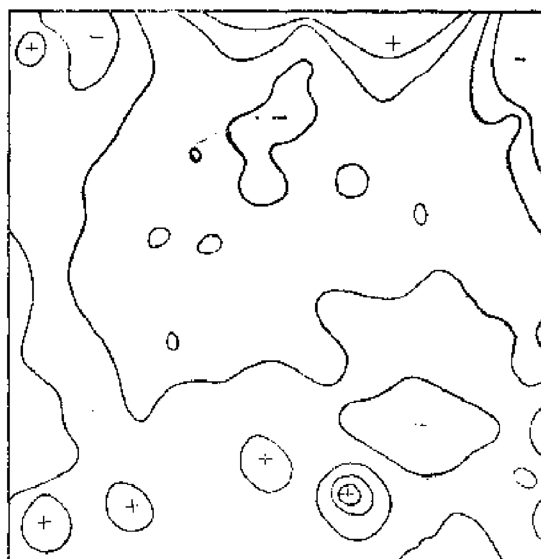


图2-2-13　五次残差图

一、傅立叶级数趋势面方程的建立

设有一组n个观测数据，观测点的平面坐标为（x_i，y_i），地质变量的观测值为z_i（$i=1, 2, \cdots, n$），为了拟合观测值z_i，可用二维傅立叶级数建立趋势面方程。其数学表达式为

$$\hat{z}=\sum_{t=0}^{r}\sum_{k=0}^{r}\Big[a_{tk}\cos\frac{2t\pi x}{L}\cos\frac{2k\pi y}{H}+b_{tk}\sin\frac{2t\pi x}{L}\cos\frac{2k\pi y}{H}$$

$$+c_{tk}\cos\frac{2t\pi x}{L}\sin\frac{2k\pi y}{H}+d_{tk}\sin\frac{2t\pi x}{L}\sin\frac{2k\pi y}{H}\Big] \qquad (2\text{-}2\text{-}7)$$

式中 $\hat{z}$——傅立叶级数趋势值；

r——x方向上给定的傅立叶级数的最大调和次数；

s——y方向上给定的傅立叶级数的最大调和次数；

a_{tk}——x方向上为t次，y方向上为k次的余弦—余弦项系数；

b_{tk}——x方向上为t次，y方向上为k次的正弦—余弦项系数；

c_{tk}——x方向上为t次，y方向上为k次的余弦—正弦项系数；

d_{tk}——x方向上为t次，y方向上为k次的正弦—正弦项系数；

L——x方向的取样长度，即原图的横向长度；

H——y方向的取样长度，即原图的纵向长度。

为了计算上的方便，傅立叶级数趋势方程一般采用其展开形式，例如，一次（即$r=1$，x方向为一次；$s=1$，y方向也为一次）傅立叶级数趋势面方程可表示为如下形式

$$\hat{z}=a_{00}+a_{10}A_1C_0+a_{01}A_0C_1+a_{11}A_1C_1+b_{10}B_1C_0+b_{11}B_1C_1+C_{01}A_0D_1+C_{11}A_1D_1+d_{11}B_1D_1 \quad (2\text{-}2\text{-}8)$$

一次傅立叶级数趋势面方程中有9个待定系数：a_{00}，a_{10}，a_{01}，a_{11}，b_{10}，b_{11}，c_{01}，c_{11}，d_{11}。

（2-2-8）式中

$$A_t=\cos\frac{2t\pi x}{L}$$

$$B_t=\sin\frac{2t\pi x}{L}$$

$$C_k=\cos\frac{2k\pi y}{H}$$

$$D_k=\sin\frac{2k\pi y}{H} \qquad (t=0,\ 1;\ k=0,\ 1)$$

为求出方程中的待定系数，可按最小二乘法原理，使每个待定系数对观测值与趋势值的总离差平方和Q的偏导数为0，即

$$Q=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2=\sum_{i=1}^{n}(z_i-a_{00}-a_{10}A_1C_0-a_{01}A_0C_1-a_{11}A_1C_1-b_{10}B_1C_0-b_{11}B_1C_1-c_{01}A_0D_1-c_{11}A_1D_1-d_{11}B_1D_1)^2$$

$$\frac{\partial Q}{\partial a_{00}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-A_0C_0)=0$$

$$\frac{\partial Q}{\partial a_{10}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-A_1C_0)=0$$

$$\frac{\partial Q}{\partial a_{01}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-A_0C_1)=0$$

$$\frac{\partial Q}{\partial a_{11}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-A_1C_1)=0$$

$$\frac{\partial Q}{\partial b_{10}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-B_1C_0)=0$$

$$\frac{\partial Q}{\partial b_{11}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-B_1C_1)=0$$

$$\frac{\partial Q}{\partial c_{01}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-A_0D_1)=0$$

$$\frac{\partial Q}{\partial c_{11}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-A_1D_1)=0$$

$$\frac{\partial Q}{\partial d_{11}}=2\sum_{i=1}^{n}(z_i-\hat{z}_i)(-B_1D_1)=0$$

上述9个方程式经整理可得到一次傅立叶级数趋势面的正规方程组，可写成如下矩阵形式

$$\begin{pmatrix} \Sigma 1 & \Sigma A_1C_0 & \Sigma A_0C_1 & \Sigma A_1C_1 & \cdots & \Sigma B_1D_1 \\ \Sigma A_1C_0 & \Sigma(A_1C_0)^2 & \Sigma A_1C_0A_0C_1 & \Sigma A_1C_0A_1C_1 & \cdots & \Sigma A_1C_0B_1D_1 \\ \Sigma A_0C_1 & \Sigma A_0C_1A_1C_0 & \Sigma(A_0C_1)^2 & \Sigma A_0C_1A_1C_1 & \cdots & \Sigma A_0C_1B_1D_1 \\ \Sigma A_1C_1 & \Sigma A_1C_1A_1C_0 & \Sigma A_1C_1A_0C_1 & \Sigma(A_1C_1)^2 & \cdots & \Sigma A_1C_1B_1D_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \Sigma B_1D_1 & \Sigma B_1D_1A_1C_0 & \Sigma B_1D_1A_0C_1 & \Sigma B_1D_1A_1C_1 & \cdots & \Sigma(B_1D_1)^2 \end{pmatrix}\begin{pmatrix} a_{00} \\ a_{10} \\ a_{01} \\ a_{11} \\ \cdots \\ d_{11} \end{pmatrix}=\begin{pmatrix} \Sigma z \\ \Sigma zA_1C_0 \\ \Sigma zA_0C_1 \\ \Sigma zA_1C_1 \\ \cdots \\ \Sigma zB_1D_1 \end{pmatrix}$$

按类似方法，可以得到x方向为r次，y方向为s次的高次调和趋势面的正规方程组。

傅立叶级数趋势面方程的拟合度计算、残差分析方法与多项式趋势面完全相同，在此就不一一叙述了。

二、调合趋势面方程次数与待定系数个数之间的关系

由（2-2-7）式得知调和趋势面方程的值是由傅立叶级数两次求和得到的，即趋势值$\hat{z}$与x方向的调和次数t及y方向的调和次数k都有关，而求和时的调和次数都是从0开始的，当$t=0$，$k=0$时，因$\sin 0=0$，所以

$$B_t=\sin\frac{2t\pi x}{L}=0$$

$$D_k=\sin\frac{2k\pi y}{H}=0$$

这就使得$B_0C_0=A_0D_0=B_0D_0=0$。而$t=0$，$k=0$时，因$\cos 0=1$，所以

$$A_t=\cos\frac{2t\pi x}{L}=1$$

$$C_k=\cos\frac{2k\pi y}{H}=1$$

这就使得$A_0C_0=1$。

可见，在调和趋势方程中A_tC_k项、B_tC_k或A_tD_k项、B_tD_k项的个数是不相同，这就给

建立调和趋势面方程的系数矩阵带来一些麻烦。0至6次调和趋势面方程的 A_iC_k、B_iC_k、A_iD_k、B_iD_k 各项待定系数的个数详见表2-2-2。

表2-2-2　调和趋势面方程的次数与待定系数的个数关系表

乘积项 \ 方程次数	0	1	2	3	4	5	6
AC	1(1)	3(4)	5(9)	7(16)	9(25)	11(36)	13(49)
BC	0(0)	2(2)	4(6)	6(12)	8(20)	10(30)	12(42)
AD	0(0)	2(2)	4(6)	6(12)	8(20)	10(30)	12(42)
BD	0(0)	1(1)	3(4)	5(9)	7(16)	9(25)	11(36)
总数	1(1)	8(9)	16(25)	24(49)	32(81)	40(121)	48(169)

表2-2-2中的不带括号的数字是对应于方程次数所新增加的待定系数个数；而括号中的数字是对应于方程次数的累计待定系数个数。

调和趋势面方程待定系数的个数由4部分组成，即x方向为r次、y方向为s次调和趋势面方程待定系数的个数为

$$
\begin{aligned}
S &= S_a + S_b + S_c + S_d \\
&= (r+1)(s+1) + r(s+1) + (r+1)s + rs \\
&= 4rs + 2r + 2s + 1
\end{aligned}
$$

当$r=s$时，即x方向与y方向的调和次数相等时则有

$$
\begin{aligned}
S &= S_a + S_b + S_c + S_d \\
&= (r+1)^2 + r(r+1) + (r+1)r + r^2 \\
&= 4r^2 + 4r + 1
\end{aligned}
$$

三、算　例

某探区的地质构造具有明显的波状起伏特点，因而采用调合趋势分析方法进行研究。表2-2-3中给出了某个标准层90个观测点的横坐标x、纵坐标y以及高程值z，标准层的高程是由地震资料解释得到的。表中的井位坐标是以图2-2-14的左下角为坐标原点（0，0）实际测量到的。图2-2-14是90个观测点的位置图。

表2-2-3　观测点坐标及标准层高程数据表

序号	横坐标, x	纵坐标, y	高程值, z(m)	序号	横坐标, x	纵坐标, y	高程值, z(m)
1	1.0	9.0	800.0	10	10.0	9.0	875.0
2	2.0	9.0	800.0	11	1.0	8.0	800.0
3	3.0	9.0	800.0	12	2.0	8.0	800.0
4	4.0	9.0	875.0	13	3.0	8.0	875.0
5	5.0	9.0	875.0	14	4.0	8.0	875.0
6	6.0	9.0	875.0	15	5.0	8.0	875.0
7	7.0	9.0	875.0	16	6.0	8.0	875.0
8	8.0	9.0	900.0	17	7.0	8.0	875.0
9	9.0	9.0	875.0	18	8.0	8.0	875.0

续表

序号	横坐标,x	纵坐标,y	高程值,z(m)	序号	横坐标,x	纵坐标,y	高程值,z(m)
19	9.0	8.0	875.0	55	5.0	4.0	1650.0
20	10.0	8.0	950.0	56	6.0	4.0	1625.0
21	1.0	7.0	800.0	57	7.0	4.0	1650.0
22	2.0	7.0	875.0	58	8.0	4.0	1600.0
23	3.0	7.0	875.0	59	9.0	4.0	1120.0
24	4.0	7.0	900.0	60	10.0	4.0	1475.0
25	5.0	7.0	875.0	61	1.0	3.0	1800.0
26	6.0	7.0	875.0	62	2.0	3.0	1725.0
27	7.0	7.0	950.0	63	3.0	3.0	1550.0
28	8.0	7.0	1100.0	64	4.0	3.0	1650.0
29	9.0	7.0	1350.0	65	5.0	3.0	1425.0
30	10.0	7.0	1650.0	66	6.0	3.0	1200.0
31	1.0	6.0	875.0	67	7.0	3.0	1500.0
32	2.0	6.0	880.0	68	8.0	3.0	1200.0
33	3.0	6.0	875.0	69	9.0	3.0	1000.0
34	4.0	6.0	890.0	70	10.0	3.0	1030.0
35	5.0	6.0	1050.0	71	1.0	2.0	1550.0
36	6.0	6.0	1250.0	72	2.0	2.0	1300.0
37	7.0	6.0	1500.0	73	3.0	2.0	1300.0
38	8.0	6.0	1800.0	74	4.0	2.0	1700.0
39	9.0	6.0	1750.0	75	5.0	2.0	1200.0
40	10.0	6.0	1700.0	76	6.0	2.0	1120.0
41	1.0	5.0	875.0	77	7.0	2.0	930.0
42	2.0	5.0	880.0	78	8.0	2.0	890.0
43	3.0	5.0	1050.0	79	9.0	2.0	890.0
44	4.0	5.0	1325.0	80	10.0	2.0	900.0
45	5.0	5.0	1700.0	81	1.0	1.0	1600.0
46	6.0	5.0	1800.0	82	2.0	1.0	1500.0
47	7.0	5.0	1700.0	83	3.0	1.0	1100.0
48	8.0	5.0	1450.0	84	4.0	1.0	1050.0
49	9.0	5.0	1500.0	85	5.0	1.0	900.0
50	10.0	5.0	1400.0	86	6.0	1.0	890.0
51	1.0	4.0	1040.0	87	7.0	1.0	890.0
52	2.0	4.0	1300.0	88	8.0	1.0	890.0
53	3.0	4.0	1800.0	89	9.0	1.0	900.0
54	4.0	4.0	1800.8	90	10.0	1.0	900.0

经计算x方向一次y方向一次至x方向三次y方向三次的调和趋势面图见图2-2-15至图2-2-23。

x方向一次y方向一次的调和趋势面方程为

$$\hat{z}=1140.250-243.813A_0C_1-7.160A_1C_0+139.888A_1C_1+5.827B_1C_0+215.804B_1C_1+179.688A_0D_1-58.298A_1D_1+253.935B_1D_1$$

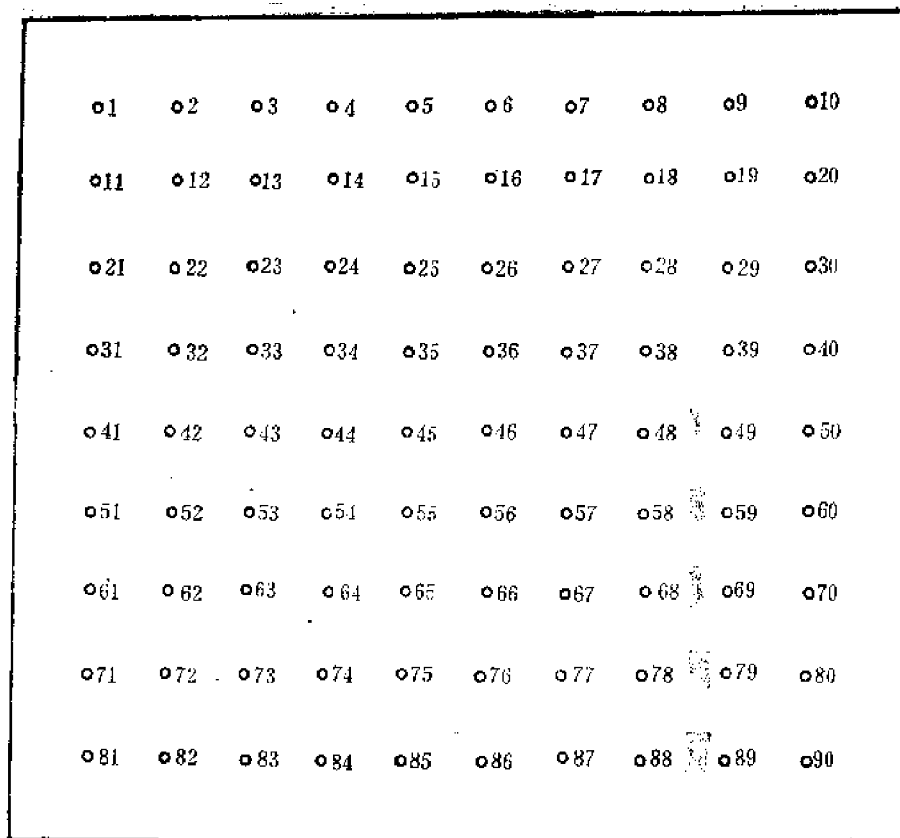


图2-2-14　观测点位置图

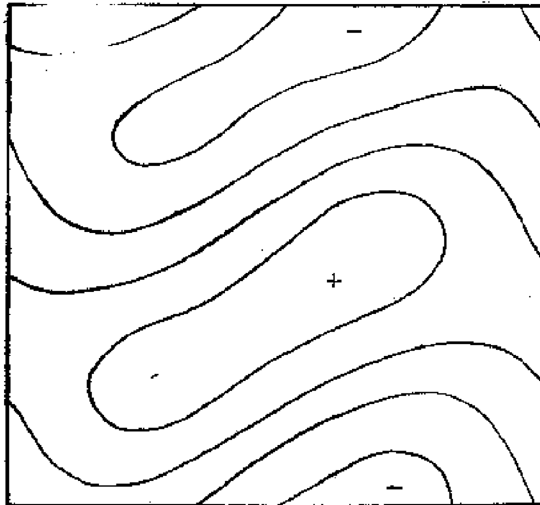


图2-2-15　x方向一次y方向一次调和趋势面图

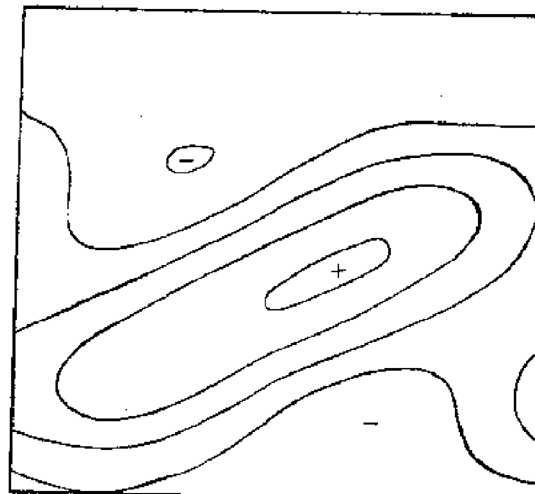


图2-2-16　x方向一次y方向二次调和趋势面图



图2-2-17 x方向一次y方向三次调和趋势面图

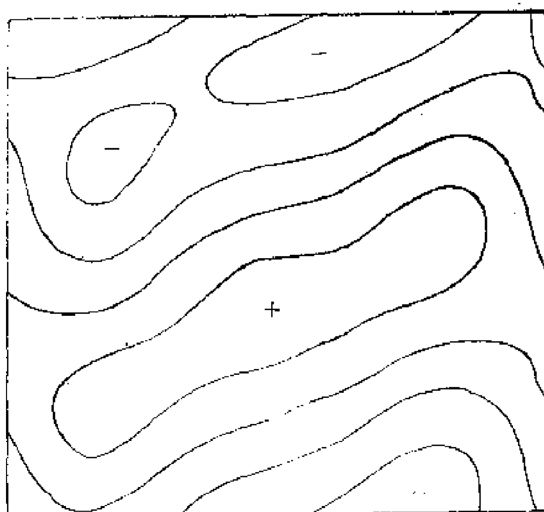


图2-2-18 x方向二次y方向一次调和趋势面图

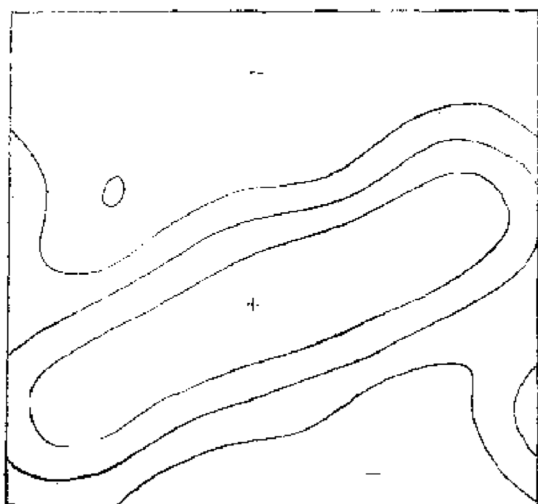


图2-2-19 x方向二次y方向二次调和趋势面图

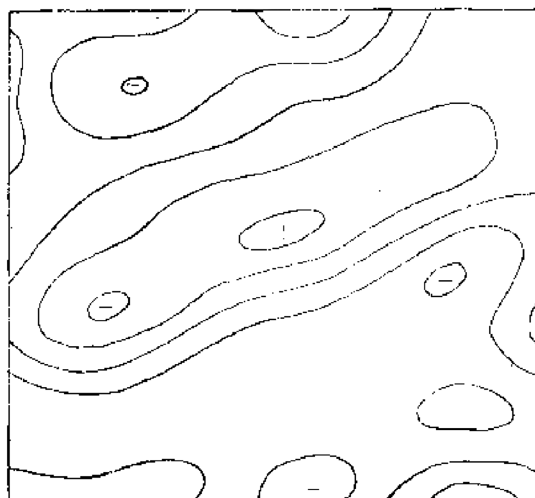


图2-2-20 x方向二次y方向三次调和趋势面图

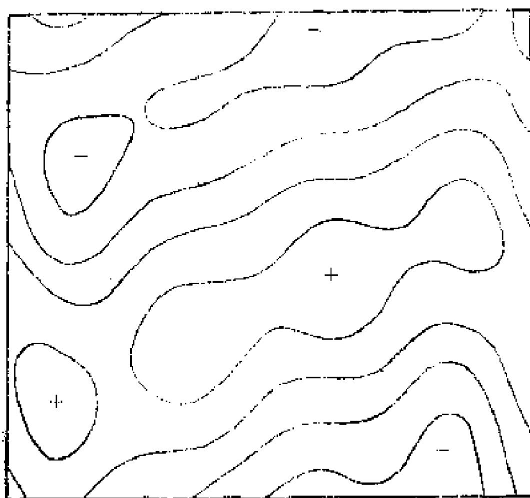


图2-2-21 x方向三次y方向一次调和趋势面图

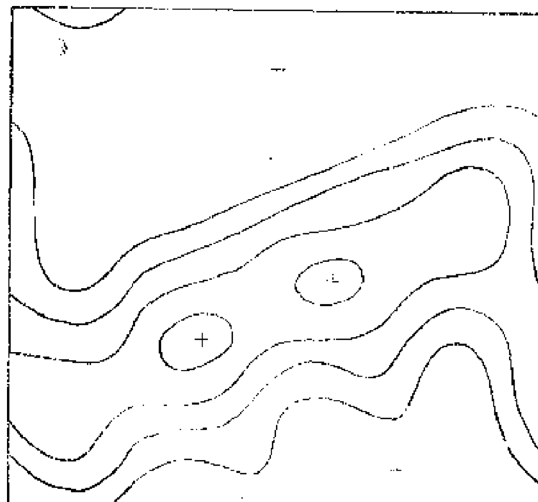


图2-2-22 x方向三次y方向二次调和趋势面图

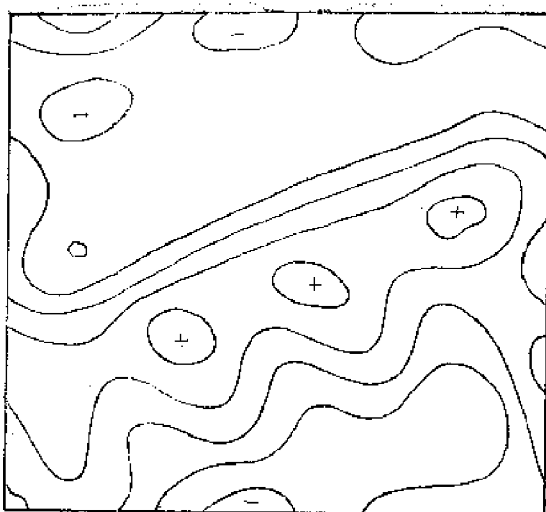


图2-2-23　x方向三次y方向三次调和趋势面图

拟合度　$C=0.730$。

x方向二次y方向二次的调和趋势面方程为

$$
\begin{aligned}
\hat{z}= & 1143.158-238.638A_0C_1 \\
& +13.727A_0C_2-17.388A_1C_0 \\
& +118.150A_1C_1-84.853A_1C_2 \\
& +9.636A_2C_0+14.142A_2C_1 \\
& +14.414A_2C_2-11.667B_1C_0 \\
& +180.816B_1C_1-122.458B_1C_2 \\
& -8.220B_2C_0+79.980B_2C_1 \\
& +10.346B_2C_2+179.489A_0D_1 \\
& -0.939A_0D_2-58.695A_1D_1 \\
& +139.234A_1D_2-1.587A_2D_1 \\
& +16.491A_2D_2+253.935B_1D_1-68.808B_1D_2+91.360B_2D_1-2.072B_2D_2
\end{aligned}
$$

拟合度　$C=0.863$

x方向三次y方向三次的调和趋势面方程为

$$
\begin{aligned}
\hat{z}= & 1157.577-224.282A_0C_1+24.966A_0C_2+47.379A_0C_3+15.979A_1C_0 \\
& +155.921A_1C_1-53.316A_1C_2+117.409A_1C_3+8.181A_2C_0-17.731A_2C_1 \\
& -23.692A_2C_2-56.701A_2C_3+32.435A_3C_0-22.025A_3C_1-40.724A_3C_2 \\
& +13.886A_3C_3-13.869B_1C_0+176.411B_1C_1-126.863B_1C_2-11.012B_1C_3 \\
& -3.388B_2C_0+89.645B_2C_1+20.010B_2C_2+24.162B_2C_3-0.878B_3C_0 \\
& +54.011B_3C_1-24.576B_3C_2-18.459B_3C_3+188.296A_0D_1-7.230A_0D_2 \\
& +5.605A_0D_3-41.083A_1D_1+126.652A_1D_2-0.181A_1D_3+16.025A_2D_1 \\
& +3.909A_2D_2+0.945A_2D_3+52.837A_3D_1-37.745A_3D_2+3.185A_3D_3 \\
& +253.935B_1D_1-68.808B_1D_2+52.437B_1D_3+91.369B_2D_1-2.072B_2D_2 \\
& -12.425B_2D_3+62.332B_3D_1+20.044B_3D_2-36.763B_3D_3
\end{aligned}
$$

拟合度　$C=0.943$

由以上计算结果可以看出，调和趋势面对波状起伏的标准层具有较好的拟合效果。随着方程次数的增高，拟合程度逐渐增高。

这里特别需要指出，调和趋势面分析方法可以进行x方向与y方向次数不相同的趋势面计算，这对于分析某一方向起伏变化的次数大于另一方向的地质界面显然是有益处的。

第三节　小　　结

（1）目前绝大多数的多元统计方法都是通用性的方法，因而，对各种学科中的统计问题都适用。但是，唯独趋势分析更适用于研究地质方面的问题，并且在实际应用中已取得良好效果。

（2）在实际使用时，确定趋势面方程的次数是一个重要问题。对于起伏不大、形态简

单的地质曲面，可以用低次趋势面拟合；而对于起伏较大、形态复杂的地质曲面，可以用高次趋势面拟合。但是，一般来说，趋势面的次数不宜过高，其原因有三个方面：一是次数较高的趋势面只在观测点附近的效果较好，而远离观测点的地方效果较差；二是对于次数较高的趋势面往往很难从地质概念上作出合理的解释；三是次数较高的趋势面，一般拟合度偏高，容易使残差信息丢失，因而不利于发现局部异常。

（3）确定趋势分析的拟合度也是一个重要问题。作为地质找矿，可根据地质条件相似的探区（也可称为模型区或演习区）所积累的经验确定。拟合度太高会使异常区消失；拟合度太低又没有完全消除区域性背景因素，也不利于选择异常区。就一般而论，拟合度以70%～90%为宜。

（4）高次趋势面可以作为拟合地质曲面的一种方法，当趋势面的拟合度大于拟合精度要求时，趋势面可作为地质曲面的近似曲面。

第三章 聚类分析

俗话说："人以群分，物以类聚"，"不知其人，观之其友"，说的都是按事物属性对其进行归类的问题。

聚类分析又称群分析、簇分析、点群分析。它是按着一批研究对象在性质上的亲疏关系进行分类的一种多元统计分析方法。聚类分析的方法很多，而且近年还在不断扩充其研究领域。聚类分析的方法类型，可按如下三种情况进行归类：

（1）按分类的对象不同，可分为两种：对样品进行分类，称为Q型聚类分析；对变量进行分类，称为R型聚类分析。

（2）按分类对象之间的关系可分为两种：当分类对象之间无次序约束关系时，称为无序量聚类分析；当分类对象之间存在次序约束关系时，称有序量聚类分析。一般情况下，有序量聚类的分类对象都是样品，而对变量之间不存在次序约束关系的，按样品聚类空间的维数又有一维有序分割法与二维（平面）有序分割法之分。

（3）按聚类分析的方法原理可分为5类：

①聚合法：分类开始时每个样品自成一类，然后按某种分类统计量使最亲近的类合并，使类的数目减少，直到所有样品合并成一类为止。分类结果常用分类谱系图（也叫聚类树）表示样品间的亲疏关系。聚合法是目前最常用的聚类分析方法。

②分裂法：分类开始时将全部样品看成一类，然后根据某种准则进行分裂，一直分裂到所需要的分类为止。是否需要分裂，通常用一个分类函数（也称误差函数或目标函数）来控制，并且规定分类合理时分类函数值最小，分类不合理时分类函数值就大。由于将n个样品（特别是当n很大时）分成k类的各种可能分法极多，以致在通常情况下无法求得精确最优解。因而，分裂法通常只能是求局部最优解的一种方法。

③调优法：开始分类时，首先对样品粗糙地分个类，然后以分类函数尽可能小为原则对分类进行调整，直到认为分类合理为止。动态聚类法就是其中最典型的方法，这种方法首先选个凝聚点，再给个初始分类，看分类是否合理，如果分类合理则计算结束；如果分类不合理则按某种原则修改分类，直到分类合理为止。

④加入法：加入法的本来含义是业已存在一个分类结果，如果又有一批新的样品需要加入到这个已存在的分类系统之中，则按某种准则确定每个新加入样品在分类结构中最合适的位置。后来，加入法演变为一种新的聚类方法，即开始时输入两个样品并看成一类，然后将其余样品逐个加入，分类结构逐步扩大，当样品全部加入后，则得到最后的分类。

⑤图论法：近年图论方法已被应用到聚类分析中，图论中的最小支撑树不仅可用于分类，还可用来定义分类函数，而且可以讨论分类函数的性质优劣。

除上述5种类型聚类分析方法外，近年来又出现了预报法及变量筛选法。预报法是指用自变量x的值来预测因变量y值的方法。预报在回归分析中已比较成熟，而聚类分析也可用作预报，特别是在回归预报中效果不好时，用聚类分析进行预报却往往能得到满意的结果。对于变量筛选，在回归分析、判别分析中早已出现筛选变量的逐步回归、逐步判别方法，而

以往聚类分析却很少考虑变量的筛选问题，最近已经出现一些尚不成熟的筛选变量的聚类方法。

第一节　分类统计量

假设有n个样品，每个样品观测了m个变量（指标）。分类统计量就是依据这些原始数据所建立的分类指标。建立分类统计量有两个途径，其一是把每个样品看作为m维空间中的一个点，在点与点之间定义某种距离系数；其二是把每个样品看作为m维空间中的一个向量，在向量与向量之间定义某种相似系数。距离系数与相似系数都可以作为衡量样品之间亲疏关系的分类统计量。如果是对变量进行分类，则应该把每个变量看作是n维空间的一个点或一个向量，同样可以定义某种距离系数或相似系数，作为衡量变量之间亲疏关系的分类统计量。

一、距离系数

如果对由n个样品，m个变量构成的研究对象进行分类时，对于每个样品可以看作是m维变量空间中的一个点，每个变量可以看作是n维样品空间中的一个点。因而，可以从几何学角度定义点与点之间的距离，不妨以d_{ij}表示x_i与x_j之间的距离。

目前，在实际工作中，最常用的是明考斯基（Minkowski）距离$d_{ij}(q)$。对于行为样品，列为变量的原始数据矩阵，样品之间的明氏距离可表示为

$$d_{ij}(q)=\left[\sum_{k=1}^{m}|x_{ik}-x_{jk}|^{q}\right]^{\frac{1}{q}} \quad (2\text{-}3\text{-}1)$$

$$(i,\ j=1,\ 2,\ \cdots,\ n)$$

而变量之间的明氏距离也可以表示为

$$d_{ij}(q)=\left[\sum_{k=1}^{n}|x_{ki}-x_{kj}|^{q}\right]^{\frac{1}{q}}$$

在明氏距离中，$q=1,\ 2,\ \infty$时用的较普遍。

$$d_{ij}(1)=\sum_{k=1}^{m}|x_{ik}-x_{jk}| \quad (2\text{-}3\text{-}2)$$

$$d_{ij}(2)=\left[\sum_{k=1}^{m}(x_{ik}-x_{jk})^{2}\right]^{\frac{1}{2}} \quad (2\text{-}3\text{-}3)$$

$$d_{ij}(\infty)=\max_{1\leq k\leq m}|x_{ik}-x_{jk}| \quad (2\text{-}3\text{-}4)$$

其中，$d_{ij}(1)$称为绝对（值）距离；$d_{ij}(2)$称为欧几里德距离；$d_{ij}(\infty)$称为车比雪夫距离。$d_{ij}(2)$简称欧氏距离，与通常习惯上的空间距离概念完全一致，所以在研究工作中用的最多，效果也最好。

但是，明氏距离有如下两个缺点：

（1）明氏距离与各变量的量纲有关，也就是说与变量的数值大小有关。为了克服这一缺点，可用第一篇第三章第一节定量数据的标准化方法进行处理，其中标准差标准化、极差标准化、极差正规化是最常用的标准化方法。

（2）明氏距离没有考虑变量之间的相关性。例如，有5个变量，其中4个变量之间有某种相关关系，或者说这4个变量只能反映某个特征F_1，而另外一个变量就能反映特征F_2。如果对5个变量同等看待，则实际上特征F_1的权值为4，特征F_2的权值为1。这显然夸大了特征F_1，贬低了特征F_2。

为了克服这一缺点，有时采用马氏（P.C.Mahalanobis）距离$d_{ij}(M)$作为分类统计量。

$$d_{ij}^2(M)=(X_i-X_j)'S^{-1}(X_i-X_j) \quad (2\text{-}3\text{-}5)$$

$$(i,\ j=1,\ 2,\ \cdots,\ n)$$

式中 X_i，X_j——分别为样品i、j的m个变量所组成的向量，即数据矩阵中第i、j个样品向量；

S^{-1}——为协方差逆矩阵。

但是，在未形成分类之前，用全部数据计算的均值和协方差矩阵来求马氏距离，其效果也并不一定理想，原因是马氏距离本身就依赖于类的划分。所以，在聚类分析中用马氏距离作为分类统计量并不多见。

鉴于此，在进行聚类分析之前，原则上应当对变量进行筛选。

二、相 似 系 数

除距离系数外，聚类分析中更常用的分类统计量是相似系数。

1. 夹角余弦

有时夹角余弦也称为相似系数。如果是对样品进行分类，可把每个样品看成是m维变量空间中的一个向量。那么，样品$X_i=(x_{i1},\ x_{i2},\ \cdots,\ x_{im})$与样品$X_j=(x_{j1},x_{j2},\cdots,x_{jm})$之间的相似程度就可用这两个向量间的夹角余弦$\cos\theta_{ij}$来表示，即

$$\cos\theta_{ij}=\frac{X_iX_j}{|X_i|\ |X_j|}=\frac{\sum_{k=1}^{m}x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{m}x_{ik}^2\cdot\sum_{k=1}^{m}x_{jk}^2}} \quad (2\text{-}3\text{-}6)$$

$$(i,\ j=1,\ 2,\ \cdots,\ n)$$

2. 相关系数

相关系数就是数理统计学中常用的皮尔逊相关系数，它是数据经过标准差标准化变换后的夹角余弦，一般用r_{ij}表示，即

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)^2\cdot\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)^2}} \quad (2\text{-}3\text{-}7)$$

相关系数是一个很重要的分类统计量，也是聚类分析中最常用的分类统计量。在第一篇第三章第五节混合数据预处理中已经谈到相关系数与距离系数可以互换的问题。由相关系数r_{ij}可以很方便地导出样品（或变量）之间的距离系数，即

$$d_{ij}^2=1-r_{ij}^2\text{或}d_{ij}=1-|r_{ij}|$$

第二节　聚合法聚类分析

聚合法是将类由多变少的一种聚类分析方法，是目前最常用的聚类分析方法。共计算过程是：

（1）开始时，每个样品自成一类，共有n类；

（2）按某种分类统计量，计算样品间的亲疏关系，将最亲近的两个样品合并成一类，形成一个由两个样品组成的样品集团（类）；

（3）计算新类与其余各类之间的亲疏关系，再将最亲近的两类合并。此时如果类的数目仍然大于1，则继续重复本步骤，直到所有类归为一类，则停止计算。

如果用距离系数研究样品分类时，如前所述，每个样品可看作m维空间中的一个点。但是，当最亲近的两个样品合并成一个类时，则形成一个样品集团，即m维空间中的一个点群。

作为单个样品之间的距离，在概念上是明确的。然而，由若干个样品结合为一个样品集团，即一个类后，类与类之间的距离在概念上就不那么单纯了。例如，可以定义类与类之间的距离为两类之间最近样品的距离；也可以定义为两类之间最远样品的距离；或者定义为两类之间的重心距离等等。因而，由于类与类之间的距离采用了不同的定义，也就产生了多种聚合聚类分析方法。

这里约定用d_{ij}表示样品x_i与x_j之间的距离，用D_{ij}表示类G_i与G_j之间的距离。

一、最短距离法

若类G_p与类G_q之间的距离D_{pq}定义为

$$D_{pq}=\min_{\substack{x_i\in G_p\\ x_j\in G_q}} d_{ij} \tag{2-3-8}$$

上式中的D_{pq}为类G_p与G_q中相距最近的两个样品的距离。用这个距离进行聚合聚类时，称最短距离法。其计算步骤可以归结为如下四步：

1.　计算所有样品与样品之间的距离

$$D^{(0)}=[d_{ij}]_{n\times n}$$

$D^{(0)}$表示每个样品各为一类的初始距离系数矩阵。此时，显然有$D_{pq}=d_{pq}$。

2.　$D^{(0)}$中的主对角线上的元素d_{ii}

表示样品x_i与自身的距离，显然$d_{ii}=0$。因而，要在$D^{(0)}$中寻找非对角线元素中距离值为最小的元素，若为d_{pq}，则将G_p与G_q合并成一个新类，记为G_r，$G_r=\{G_p,\ G_q\}$。

3.　计算新类与其他类之间的距离

$$\begin{aligned}D_{rk}&=\min_{\substack{x_i\in G_r\\ x_j\in G_k}} d_{ij}\\&=\min\{\min_{\substack{x_i\in G_p\\ x_j\in G_k}} d_{ij},\ \min_{\substack{x_i\in G_q\\ x_j\in G_k}} d_{ij}\}\\&=\min\{D_{pk},\ D_{qk}\}\end{aligned} \tag{2-3-9}$$

$(k=1, 2, \cdots, n; \ k \neq p, q)$

将$D^{(0)}$中的第p、q行及p、q列的元素删去。在此不妨约定$p<q$，保留小号p作为新类的样品号。并将刚计算的D_{rk}写在p行、p列上，所形成的距离系数矩阵记为$D^{(1)}$。

$$D^{(1)}=[D_{ij}]_{(n-1)\times(n-1)}$$

4. 对$D^{(1)}$重复进行步骤2、3

从而得到$D^{(2)}$，由$D^{(2)}$按同样步骤得到$D^{(3)}$，…，直到所有样品都合并成一类为止。

如果在某一步的$D^{(s)}$中有两个或两个以上最小值相等的元素，则这些元素可以同时归为一类。最短距离法不仅可用于样品分类，即Q型聚类；也可用于变量分类，即R型聚类。

最短距离法的分类统计量如果是相似性指标，例如$\cos\theta$或r，则（2-3-8）式及（2-3-9）式中min换成max即可。其原因是$\cos\theta_{ii}=1$，$r_{ii}=1$，即类G_i与自身的相似性指标为1。

二、最长距离法

如果类与类之间的距离定义为两类中相距最远样品的距离，即

$$D_{pq}=\max_{\substack{x_i\in G_p\\ x_j\in G_q}} d_{ij} \qquad (2\text{-}3\text{-}10)$$

则称为最长距离法。

最长距离法与最短距离法的并类步骤完全相同，也是每个样品先自成一类，然后将距离最小的两类合并为G_r，则G_r与某类G_k的距离为

$$\begin{aligned} D_{rk}&=\max_{\substack{x_i\in G_r\\ x_j\in G_k}} d_{ij}\\ &=\max\{\max_{\substack{x_i\in G_p\\ x_j\in G_k}} d_{ij}, \max_{\substack{x_i\in G_q\\ x_j\in G_k}} d_{ij}\}\\ &=\max\{D_{pk}, \ D_{qk}\} \qquad (2\text{-}3\text{-}11) \end{aligned}$$

$(k=1, 2, \cdots, n; \ k \neq p, q)$

然后再将距离最近的两类合并，直到所有的样品全部合并为一类为止。

三、中间距离法

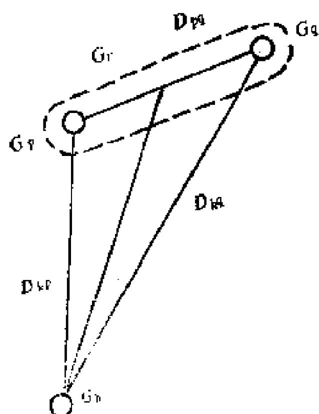


图 2-3-1 中间距离法示意图

如果类与类之间的距离既不采用最近样品之间的距离，也不采用最远样品之间的距离，而是采用最近与最远样品之间的距离进行聚类分析时，则称作中间距离法。

当聚类分析计算到某一步骤时，将G_p与G_q合并为G_r，继续计算G_r与某一类G_k（$k\neq p, q$）的距离时，为了不失一般性，这里可设$D_{kp}<D_{kq}$。按最短距离法则$D_{kr}=D_{kp}$；按最长距离法则$D_{kr}=D_{kq}$。见图2-3-1。

由图2-3-1可以看出，三角形的三个边分别是D_{kp}、D_{kq}、D_{pq}，其中D_{pq}是类G_p与G_q之间的距离，

它们已合并为新类G_r，所以用虚线圈起。长度介于D_{kp}、D_{kq}之间的线，直观上以D_{pq}边的中线为最合适，这个中线的长度等于

$$\left(\frac{1}{2}D_{kp}^2+\frac{1}{2}D_{kq}^2-\frac{1}{4}D_{pq}^2\right)^{\frac{1}{2}}$$

中间距离法就是用这个中线作为D_{kr}，即

$$D_{kr}^2=\frac{1}{2}D_{kp}^2+\frac{1}{2}D_{kq}^2-\frac{1}{4}D_{pq}^2 \qquad (2\text{-}3\text{-}12)$$

（2-3-12）式中的各项都是距离的平方项，为了计算上的方便，可将样品间的初始距离系数矩阵$D^{(0)}$中的元素一律改为d_{ij}^2，第S步的$D^{(S)}$中的元素改为D_{ij}^2。

中间距离法还可以推广到更一般的形式，把（2-3-12）式中的第三项系数用β代替，即

$$D_{kr}^2=\frac{1}{2}D_{kp}^2+\frac{1}{2}D_{kq}^2+\beta D_{pq}^2 \qquad (2\text{-}3\text{-}13)$$

（2-3-13）式中$-\frac{1}{4}\leqslant\beta\leqslant0$。

最短距离法、最长距离法都是按着类与类之间的距离定义的，任何两类之间的距离都可按（2-3-9）或(2-3-11)式进行计算，并且计算结果是唯一的，因而与类的形成过程无关。但是，中间距离法计算类与类之间的距离却与类的形成过程有关，这是中间距离法的一个缺点。

四、重 心 法

从物理学的观点来看，一个类是高维空间中的一个点群，显然用它的重心来代表这个类是合理的，如果类与类之间的距离用重心之间的距离来表示则为重心法。设类G_p、G_q的重心分别是$\bar{x}_p$、$\bar{x}_q$，则G_p与G_q之间的距离是

$$D_{pq}=d_{\bar{x}_p\bar{x}_q} \qquad (2\text{-}3\text{-}14)$$

如果某一步将G_p与G_q合并为G_r，它们的样品数分别为n_p、n_q、$n_r=n_p+n_q$，重心分别是$\bar{x}_p$、$\bar{x}_q$、$\bar{x}_r$，这些重心都是m维空间中的向量。G_r的重心为

$$\bar{x}_r=\frac{1}{n_r}(n_p\bar{x}_p+n_q\bar{x}_q) \qquad (2\text{-}3\text{-}15)$$

设某个类G_k的重心是$\bar{x}_k$，则G_k与G_r之间的距离为

$$D_{kr}^2=d^2_{\bar{x}_k\bar{x}_r}=\frac{n_p}{n_r}D_{kp}^2+\frac{n_q}{n_r}D_{kq}^2-\frac{n_p}{n_r}\cdot\frac{n_q}{n_r}D_{pq}^2 \qquad (2\text{-}3\text{-}16)$$

（2-3-16）式就是重心法计算类与类之间距离的递推公式。公式中的各项都是距离的平方，因此与中间距离法一样，将距离系数矩阵换为距离的平方矩阵更为方便。

从物理学观点看，用重心作为空间点群的代表点是比较合理的。但是，重心法不能代表类的一切特征，如图2-3-2中有两个类G_1和G_2，如果图2-3-2（a）中的重心不动将G_1转90°，得到图2-3-2（b），从直观上看（a）与（b）的相关关系并不相同，但按重心法聚类时（a）、（b）却完全一样。为了克服这一缺点，又产生了类平均法。

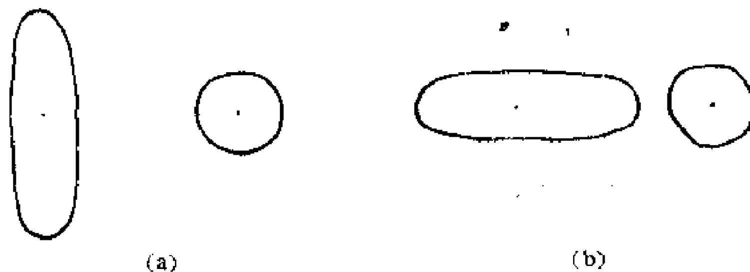


图 2-3-2 重心法不能代表类的一切特征

五、类 平 均 法

类平均法定义两类的距离平方等于两类中两两元素之间的平均平方距离，即

$$D_{pq}^2=\frac{1}{n_p n_q}\sum_{\substack{x_i\in G_p\\x_j\in G_q}} d_{ij}^2 \qquad (2\text{-}3\text{-}17)$$

上式中的n_p、n_q分别为类G_p、G_q的样品数。用这种距离作为分类统计量的聚类分析方法叫类平均法。

G_p与G_q结合为G_r后，G_r与另外的类G_k之间距离为D_{kr}^2，其递推公式为

$$D_{kr}^2=\frac{1}{n_k n_r}\sum_{\substack{x_i\in G_k\\x_j\in G_r}} d_{ij}^2=\frac{1}{n_k n_r}\Big(\sum_{\substack{x_i\in G_k\\x_j\in G_p}} d_{ij}^2+\sum_{\substack{x_i\in G_k\\x_j\in G_q}} d_{ij}^2\Big)$$

$$=\frac{n_p}{n_r}D_{kp}^2+\frac{n_q}{n_r}D_{kq}^2 \qquad (2\text{-}3\text{-}18)$$

上式中n_p、n_q分别为类G_p、G_q的样品数，$n_r=n_p+n_q$为类G_r的样品数。公式中的D_{ij}并非一定要用欧氏距离，而是任何距离都可以。所以类平均法比重心法的适应性更为广泛。

类平均法也可以用距离的加权值，即

$$D_{pq}=\frac{1}{n_p n_q}\sum_{\substack{x_i\in G_p\\x_j\in G_q}} d_{ij} \qquad (2\text{-}3\text{-}19)$$

类G_p与G_q结合为G_r后，G_r与另外的类G_k之间的距离为D_{kr}，其递推公式为

$$D_{kr}=\frac{n_p}{n_r}D_{kp}+\frac{n_q}{n_r}D_{kq} \qquad (2\text{-}3\text{-}20)$$

用这种距离定义的聚合聚类法称作加权平均法。

如果忽略n_p与n_q的差异，即认为$n_p=n_q$时，可以有更为简单的距离定义

$$D_{kr}=\frac{1}{2}(D_{kp}+D_{kq}) \qquad (2\text{-}3\text{-}21)$$

由（2-3-21）式定义距离的聚合聚类法叫作平均距离法。

有人将类平均法和中间距离法作了扩张推广。类平均法的递推公式(2-3-18)式被推广为

$$D_{kr}^2=\frac{n_p}{n_r}(1-\beta)D_{kp}^2+\frac{n_q}{n_r}(1-\beta)D_{kq}^2+\beta D_{pq}^2 \qquad (2\text{-}3\text{-}22)$$

（2-3-22）式中的$\beta<1$。用这个公式定义的聚合聚类法称为可变类平均法。

中间距离法的递推公式（2-3-13）式被推广为

$$D_{kr}^2=\frac{1-\beta}{2}(D_{kp}^2+D_{kq}^2)+\beta D_{pq}^2 \quad (2-3-23)$$

（2-3-23）式中的$\beta<1$。用这个公式定义的聚合聚类法称为可变距离法。

可变类平均法与可变距离法的分类效果与β值的关系极大，通常β以选择负值为宜。

六、聚合法小结

上述介绍的各种聚合聚类方法是目前最常用的聚类分析方法。这些方法在计算步骤上是完全一样的，所不同的仅仅是类与类之间的距离定义不同，从而导出不同的距离递推公式。

由于这些公式在形式上的不同，给研制包括各种方法的统一程序带来很多不方便。能否将这些公式表示为统一的形式呢？Wishart在1969年把这些公式统一为

$$D_{kr}^2=\alpha_p D_{kr}^2+\alpha_q D_{kp}^2+\beta D_{pq}^2+\gamma|D_{kp}^2-D_{kq}^2| \quad (2-3-24)$$

（2-3-24）式中有4个可供选择的参数，表2-3-1中给出了7种聚合聚类法的α_p、α_q、β、γ的取值。

表 2-3-1 七种聚合聚类法的参数表

方法名称	α_p	α_q	β	γ	注
最短距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	
最长距离法	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	
中间距离法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}\leqslant\beta\leqslant 0$	0	
重心法	n_p/n_r	n_q/n_r	$-n_p n_q/n_r^2$	0	必须用欧氏距离
类平均法	n_p/n_r	n_q/n_r	0	0	
可变类平均法	$(1-\beta)\frac{n_p}{n_r}$	$(1-\beta)\frac{n_q}{n_r}$	<1	0	
可变距离法	$(1-\beta)/2$	$(1-\beta)/2$	<1	0	

类似地，也可以将最短距离法、最长距离法、平均距离法、加权平均法统一为

$$D_{kr}=\alpha_p D_{kp}+\alpha_q D_{kq}+\gamma|D_{kp}-D_{kq}| \quad (2-3-25)$$

（2-3-25）式中的3个参数列于表2-3-2中。

表 2-3-2 四种聚合聚类法的参数表

方法名称	α_p	α_q	γ
最短距离法	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$
最长距离法	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
平均距离法	$\frac{1}{2}$	$\frac{1}{2}$	0
加权平均法	n_p/n_r	n_q/n_r	0

以上九种聚合聚类方法，对同一问题的计算结果并不完全一致，那么哪一种方法更好呢？目前尚无一个合适的衡量标准。

七、算　例

[1] 对南堡凹陷高尚堡地区16口井的17个样品的高压物性数据进行Q型聚类分析，选择9个测试项目作为地质变量。这9个地质变量是：地层压力系数（x_1），饱和压力（x_2），地层温度梯度（x_3），原始油气比（x_5），原油密度（x_5），原油粘度（x_6），体积系数（x_7），压缩系数（x_8），溶解系数（x_9）。

表 2-3-3　17个样品的井号层位及取样深度

样品序号	井　号	层　位	取样深度（m）	样品序号	井　号	层　位	取样深度（m）
1	高16	ES_3	3769.5	10	高36	ES_1下	2641
2	高36	Nm_{12-11}	1815.3	11	高17	ES_3	3651.4
3	庙 8	Ed_1	2290.7	12	柳12	ES_{3-5}	3229.7
4	柳 1	ES_{3-4}	3678.45	13	柳10	ES_{3-5}	3278.5
5	高43-1	Ed_2	3043.5	14	高 9	ES_1下	2934
6	高50	ES_1下	3041.9	15	高 5	ES_{3-3}	3138
7	高57	ES_1上	2466.3	16	高 2	ES_{3-3}	3201.8
8	高31	ES_{3-4}	3289.7	17	高12	ES_{3-3}	3585.7
9	高13	ES_3	3480				

17个样品的井号、取样层位、取样深度见表2-3-3。17个样品的9个变量的原始数据表见表2-3-4。

表 2-3-4　17个样品9个变量的原始数据表

样品序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
1	0.873	118	0.318	75	0.733	1.40	1.270	9.32×10^{-5}	0.542
2	1.000	142	0.336	58	0.818	2.97	1.153	9×10^{-5}	0.366
3	0.908	68	0.371	25	0.823	3.73	1.111	8.28×10^{-5}	0.367
4	0.927	55	0.349	35	0.785	2.89	1.160	93.2×10^{-5}	0.545
5	0.902	126	0.329	64	0.745	1.45	1.195	10.92×10^{-5}	0.429
6	1.071	194	0.325	172	0.653	0.70	1.145	14.89×10^{-5}	0.722
7	0.760	159	0.316	107	0.710	1.00	1.262	11.96×10^{-5}	0.553
8	1.230	62	0.343	38	0.780	2.67	1.157	8.5×10^{-5}	0.532
9	1.038	285	0.373	300	0.576	0.33	1.877	24.35×10^{-5}	0.681
10	1.311	99	0.310	66	0.745	1.39	1.277	9.59×10^{-5}	0.506
11	1.005	105	0.279	73	0.740	1.76	1.241	10.54×10^{-5}	0.600
12	0.918	145	0.375	104	0.697	1.25	1.327	15.71×10^{-5}	0.593
13	1.080	137	0.351	97	0.716	0.94	1.271	11.1×10^{-5}	0.584
14	1.196	173	0.344	133	0.684	1.10	1.373	11.67×10^{-5}	0.636
15	1.006	48	0.323	33	0.778	2.310	1.150	8.6×10^{-5}	0.583
16	1.218	74	0.350	42	0.784	5.24	1.163	12.89×10^{-5}	0.500
17	1.434	205	0.307	153	0.683	2.26	1.445	14.13×10^{-5}	0.639

对于表2-3-4中的原始数据，按极差正规化方法进行变换，以夹角余弦作为分类统计量，用加权平均法进行归类，经过计算得到如下结果，见表2-3-5及图2-3-3。这里对表2-3-5中“结合类（样品或样品集团）号”一项需要作如下说明：当类为单一样品时，类号就是样品号；当类为样品集团时，类号是指样品集团中的最小样品号。

表 2-3-5　结合类的相似系数（夹角余弦）

结合类号	cosθ	结合类号	cosθ
9,6	0.8857	5,1	0.6580
11,10	0.8848	15,2	0.5821
7,1	0.8493	17,6	0.5537
14,6	0.8295	10,1	0.5158
3,2	0.7850	12,6	0.3647
16,8	0.7829	4,2	0.3177
13,12	0.6762	2,1	−0.1485
8,2	0.6723	6,1	−1.0000

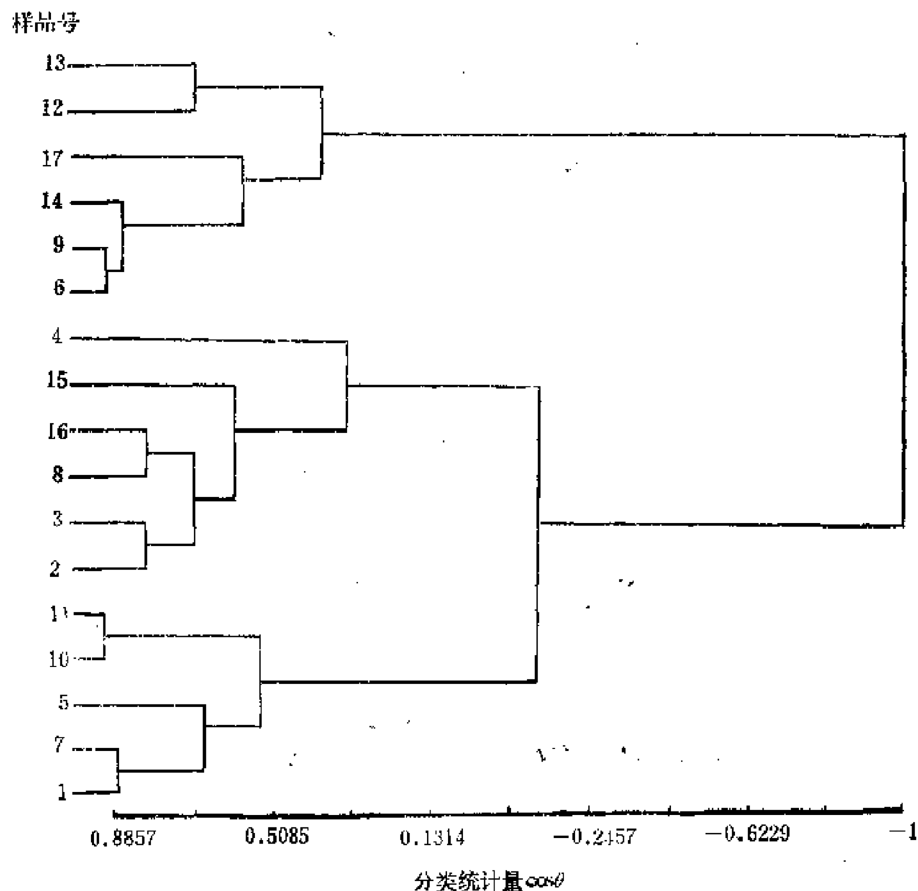


图 2-3-3　17个样品聚类谱系图

[2] 对算例1的9个变量进行R型聚类分析，采用极差正规化方法处理原始数据，以相关系数作为分类统计量，经过计算得到如下结果，见表2-3-6及图2-3-4。

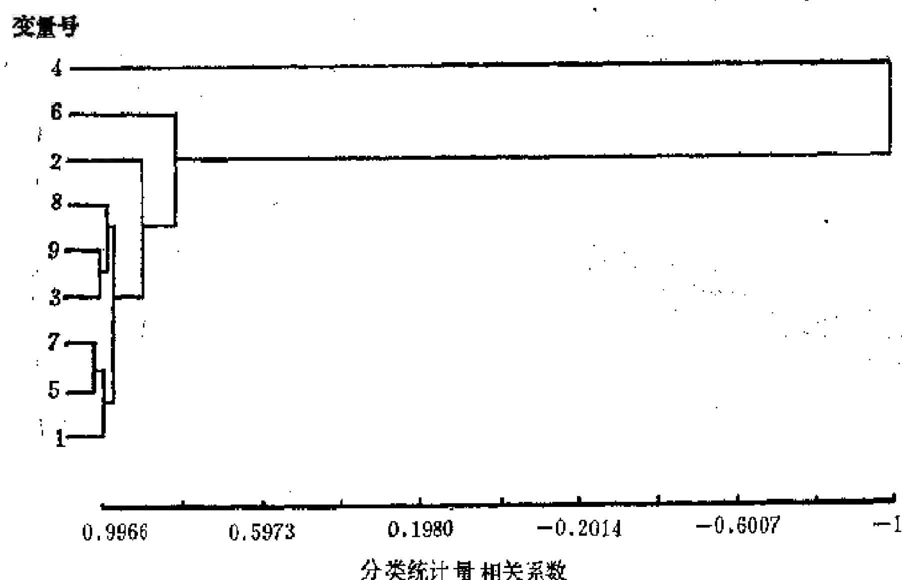


图 2-3-4 9个变量聚类谱系图

通过以上两个算例可以看出，17个高压物性样品可分为三大类，三类之间有明显区别，9个变量的变化范围及平均值列于表2-3-7中。

表 2-3-6 结合类的相关系数

结合类号	相关系数	结合类号	相关系数
7,5	0.9966	3,1	0.9571
9,3	0.9966	2,1	0.8909
5,1	0.9954	6,1	0.7996
8,3	0.9881	4,1	−1.0000

表 2-3-7 三类样品的变量统计值

变量	Ⅰ类		Ⅱ类		Ⅲ类	
	范围值	平均值	范围值	平均值	范围值	平均值
压力系数x_1	0.9177～1.4335	1.1226	0.9084～1.2299	1.0582	0.7602～1.0109	0.9108
饱和压力x_2	137～285	189.83	48～142	73.5	99～159	121.4
地温梯度x_3	0.3067～0.3507	0.3458	0.3226～0.3710	0.3453	0.2793～0.3285	0.3105
原始油气比x_4	97～300	203	25～58	38.5	64～107	77
原油密度x_5	0.5759～0.7164	0.6684	0.7777～0.8227	0.7945	0.7103～0.7450	0.7345
原油粘度x_6	0.33～2.26	1.11	2.31～5.24	3.32	1～1.76	1.4
体积系数x_7	1.271～1.877	1.4574	1.1626～1.1111	1.1489	1.2268～1.2705	1.2390
压缩系数x_8	11.1～24.35×10^{-5}	15.31×10^{-5}	8.28～12.89×10^{-5}	9.34×10^{-5}	9.32～18.92×10^{-5}	12.06×10^{-5}
溶解系数x_9	0.5839～0.8807	0.6756	0.3662～0.5833	0.4823	0.4280～0.6000	0.5380

这三类样品中，第Ⅰ类样品的压力系数（x_1）、饱和压力（x_2）、地温梯度（x_3）、原始油气比（x_4）、体积系数（x_7）、压缩系数（x_8）、溶解系数（x_9）是三类中最大的，而原油密度（x_5）、原油粘度（x_6）是三类中最小的，可见其原油性质最好，而这6口井的

产油量也是比较高的。

由R型聚类分析看出，9个变量中，压缩系数（x_8）、溶解系数（x_9）、地温梯度(x_3)聚为一类；体积系数（x_7）、原油密度（x_5）、压力系数（x_1）聚为一类，这是因为溶解系数（x_9）、压缩系数（x_8）与地温梯度（x_3）直接相关，而体积系数（x_7）、原油密度（x_5）、与压力系数（x_1）直接相关。

第三节 有序量聚类分析

有序量聚类分析是分裂法聚类分析中的一种。分裂法又称分解法，该方法在开始时将全体样品看成一类，然后将类由少变多，一直分裂到所需要的分类为止。这种分类过程与聚合聚类过程恰好相反。如何将n个样品分成两类，最基本的方法是比较一切可能的分法，选择使分类函数达到极小的一种。将n个样品分成两类的一切可能分法有$2^{(n-1)}-1$种，当n很大时计算工作量将非常大。为了节省工作量，在一般情况下不得不放弃寻求精确最优解，转而寻求局部最优解。这是分裂聚类法的不足之处。

但是，在许多地质问题中，有些地质量之间的次序是不能打乱的，例如，地层剖面中的层位由老到新的顺序是不能改变的，这种不能打乱次序的量通称有序量。显然，有序量只能是样品，而不可能是变量，变量之间有时虽然存在相关关系，然而，全部变量之间不可能构成有序约束关系。所以，严格地讲有序量聚类应称为有序样品聚类。

由于样品之间存在着次序约束关系，这就使得分裂过程的计算工作量可以大大减少。对于一维有序样品，将n个样品分成两类的一切可能分法只有（$n-1$)种，这就使得有序量聚类可以得到精确的最优解。因此，经常把一维有序样品聚类称作最优分割法。

一、最优分段的含义

有n个样品X_1，X_2，…，X_n，如果要求同类的样品必须是相互邻接的，那么，其中的某一类一定呈现如下形式

$$\{x_i,\ x_{i+1},\ \cdots,\ x_j\}$$

其中$1 \leqslant i \leqslant n$，$j \leqslant n$。

将n个样品分成k类的一切可能分法的数目为C_{n-1}^{k-1}。现在要求在所有可能的分割中找出一种分割，使得各段内部样品之间的差异最小，而各段之间的差异最大，这样的一种分割方法称为最优分割法。

所谓各段内部的差异最小，就是指各段内部表示样品特征的变量数值变化最小。在此可以用变差来表示各段内部的差异大小。例如，样品段$\{x_i,\ x_{i+1},\ \cdots,\ x_j\}$的变差可以表示为

$$d_{ij}=\sum_{k=i}^{j}[x_k-\bar{x}(i,\ j)]^2 \qquad (2\text{-}3\text{-}26)$$

式中

$$\bar{x}(i,\ j)=\frac{1}{j-i+1}\sum_{k=i}^{j}x_k \qquad (2\text{-}3\text{-}27)$$

因而，可用d_{ij}表示样品段$\{x_i, x_{i+1}, \cdots, x_j\}$内部样品间的差异情况，$d_{ij}$越小表示段内各样品之间的差异较小；反之，则表示段内各样品之间的差异较大。若是所有各段内部的差异都达到最小，则所分各段的变差总和（总变差）也就最小。

如果划分n个样品为p个样品段，则有

第一段 $\{x_{11}, x_{21}, \cdots, x_{n_1 1}\}$, $\qquad \bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{i1}$

第二段 $\{x_{12}, x_{22}, \cdots, x_{n_2 2}\}$, $\qquad \bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{i2}$

… … … … … …

第p段 $\{x_{1p}, x_{2p}, \cdots, x_{n_p p}\}$, $\qquad \bar{x}_p = \frac{1}{n_p}\sum_{i=1}^{n_p} x_{ip}$

而且有

$$\sum_{k=1}^{p} n_k = n$$

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{p}\sum_{i=1}^{n_p} x_{ik}$$

那么，总变差（离差平方和）为

$$
\begin{aligned}
S &= \sum_{k=1}^{p}\sum_{i=1}^{n_k}(x_{ik}-\bar{x})^2 \\
&= \sum_{k=1}^{p}\sum_{i=1}^{n_k}[(x_{ik}-\bar{x}_k)+(\bar{x}_k-\bar{x})]^2 \\
&= \sum_{k=1}^{p}\sum_{i=1}^{n_p}(x_{ik}-\bar{x}_k)^2+\sum_{k=1}^{p}n_k(\bar{x}_k-\bar{x})^2 \\
&\quad +2\sum_{k=1}^{p}\sum_{i=1}^{n_p}(x_{ik}-\bar{x}_k)(\bar{x}_k-\bar{x}) \\
&= \sum_{k=1}^{p}\sum_{i=1}^{n_p}(x_{ik}-\bar{x}_k)^2+\sum_{k=1}^{p}n_k(\bar{x}_k-\bar{x})^2
\end{aligned}
$$

其中

$$
\begin{aligned}
&2\sum_{k=1}^{p}\sum_{i=1}^{n_p}(x_{ik}-\bar{x}_k)(\bar{x}_k-\bar{x}) \\
&=2\sum_{k=1}^{p}(\bar{x}_k-\bar{x})\sum_{i=1}^{n_p}(x_{ik}-\bar{x}_k) \\
&=2\sum_{k=1}^{p}(\bar{x}_k-\bar{x})(n_k\bar{x}_k-n_k\bar{x}_k) \\
&=0
\end{aligned}
$$

令 $S_1 = \sum_{k=1}^{p}\sum_{i=1}^{n_k}(x_{ik}-\bar{x}_k)^2$ 为段内变差；

$$S_2=\sum_{k=1}^{p} n_k(\bar{x}_k-\bar{x})^2 \text{为段间变差。}$$

因此 $$S=S_1+S_2$$

或者 $$S_2=S-S_1$$

对于给定的n个样品，S是个确定值。若段内变差最小，则段间变差必然最大。所以，段内变差为最小的分割法就是最优分割法。

二、最优分段的计算过程

如果有n个有序样品，每个样仅有一个观测值时，则有原始数据矩阵

$$X=[x_1 x_2 \cdots x_n]$$

1. 最优二分割

对于n个有序样品进行二分割的分法，共有（$n-1$）种，即

$$\{x_1\}\{x_2, x_3, \cdots, x_n\}$$
$$\{x_1, x_2\}\{x_3, x_4, \cdots, x_n\}$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$
$$\{x_1, x_2, \cdots, x_{n-1}\}\{x_n\}$$

现在的问题是在这（$n-1$）种分法中，哪一种分法最优。为了回答这一问题需要计算这（$n-1$）种分法的总变差，其中总变差最小的那种就是最优二分法。

这里记$S_n(2, j)$为n个样品在第j点进行二分割的总变差。其中n表示被分割的样品数，2表示分成两段，j表示以第j个样品为分割点。那么，（$n-1$）种分法的总变为

$$S_n(2, 1)=d_{11}+d_{2n}$$
$$S_n(2, 2)=d_{12}+d_{3n}$$
$$\cdots \quad \cdots \quad \cdots \quad \cdots$$
$$S_n(2, n-1)=d_{1, n-1}+d_{nn}$$

其中的

$$d_{11}=d_{22}=\cdots=d_{nn}=0$$

如果$j=i1$时，$S_n(2, j)$达到最小值，即

$$S_n(2, i1)=\min_{1\leq j\leq n-1} S_n(2, j)$$

则最优二分割为

$$\{x_1, x_2, \cdots, x_{i1}\}\{x_{i1+1}, \cdots, x_n\}$$

2. 最优三分割

三分割的总变差可以记为$S_n(3, i1, j)$，其中的3表示分成三段，$i1$、j分别表示两个分割点，$2\leqslant j\leqslant n-1$，$1\leqslant i1\leqslant j-1$。

$$S_n(3, i1, j)=d_{1, i1}+d_{i1+1, j}+d_{j+1, n}$$
$$=S_j(2, i1)+d_{j+1, n}$$

如果$S_n(3, i1, j)$为最优三分割，则$S_j(2, i1)$必为最优二分割。否则必然存在另外一个最优二分割$S_j(2, i1')$，而有

$$S_n(3, i1, j)>S_n(3, i1', j)$$

这就与$S_n(3, t1, j)$为最优三分割有矛盾。可见，计算n个样品的最优三分割之前，必须先求出前$j(j=n-1, n-2, \cdots, 3, 2)$个样品的最优二分割的分割点$t1(j)$，而得到

$$\{x_1, x_2, \cdots, x_{t1(j)}\}\{x_{t1(j)+1}, \cdots, x_j\}$$

以及$\{x_{j+1}, \cdots, x_n\}$构成一个三分割，然后找出一个最好的分割点j，使得

$$S_n(3, t1(j), j)=S_j(2, t1(j))+d_{(j+1)}.$$

在所有的三分割中尽可能的小，如果$j=t2$时$S_n(3, t1(j), j)$达到最小值，即

$$S_n(3, t1, t2)=\min_{2\leqslant j\leqslant n-1} S_n(3, t1(j), j)$$

最后，得到一个最优三分割，即

$$\{x_1, x_2, \cdots, x_{t1}\}\{x_{t1+1}, \cdots, x_{t2}\}\{x_{t2+1}, \cdots, x_n\}$$

3. 最优K分割

为找出最优k分割，可以先找出前j个样品的最优$k-1$分割的总变差

$$S_j((k-1), t1(j), t2(j), \cdots, t(k-2)(j))$$
$$(j=n-1, n-2, \cdots, k-1)$$

式中的$ti(j)(i=1, 2, \cdots, k-1)$表示前$j$个数的第$i$个分割点。即有

$\{x_1, \cdots, x_{t1(j)}\}\{x_{t1(j)+1}, \cdots, x_{t2(j)}\}\cdots\{x_{t(k-2)(j)+1}, \cdots, x_j\}$与$\{x_{j+1}, \cdots, x_n\}$构成一个$k$分割，为使其是最优$k$分割，应使总变差

$$S_n(k, t1(j), t2(j), \cdots, t(k-2)(j), j)$$
$$=S_j((k-1), t1(j), \cdots, t(k-2)(j))+d_{(j+1)}.$$

为最小，而得到分割点j。设$j=t(k-1)$时，总变差$S_n(k, t1(j), \cdots, t(k-2)(j), j)$为最小，即

$$S_n(k, t1, \cdots, t(k-2), t(k-1))$$
$$=\min_{k-1\leqslant j\leqslant n-1} S_n(k, t1(j), \cdots, t(k-2)(j), j)$$

则可得最优k分割

$$\{x_1, x_2, \cdots, x_{t1}\}\{x_{t1+1}, \cdots, x_{t2}\}\cdots\{x_{t(k-1)+1}, \cdots, x_n\}$$

4. 确定分割数K

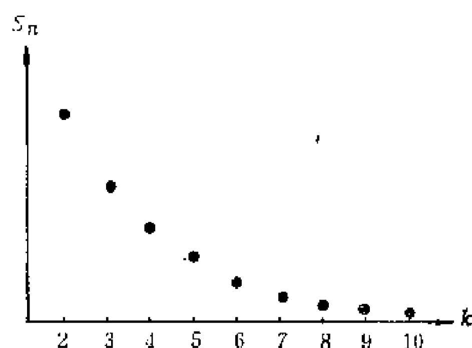


图 2-3-5 总变差与分割数的关系示意图

确定分割数k，一般有两种情况。一种情况是分割数k由实际地质情况确定，例如，我国南方的三叠系，必然要三分为上、中、下三个统；另一种情况是事先不知道分割数，此时，可以预先确定一个小的正数δ，使得k分割的总变差$S_n(k, t1, \cdots, t(k-1))<\delta$后，就不再继续分割了。为此，可以作总变差$S_n$与分割数$k$之间的关系曲线，见图2-3-5。

由图2-3-5可见，随着分割数k的增加，S_n逐渐趋于平缓，因此，可以选择开始趋于平缓的分割数为最优分割数。

5. 多个变量的最优分割

如果每个样品有m个变量，则n个样品的原始数据矩阵为

$$X=[x_{ij}]_{m\times n}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

这时，样品段$\{x_i, x_{i+1}, \cdots, x_j\}$的变差应为

$$d_{ij}=\sum_{b=i}^{j}\sum_{a=1}^{m}(x_{ab}-\bar{x}_a(i, j))^2$$

$$(i, j=1, 2, \cdots, n)$$

其中的

$$\bar{x}_a(i, j)=\sum_{b=i}^{j}x_{ab}/(j-i+1)$$

而最优分割的计算方法与一个变量时计算方法大体类似。

三、最优分割的具体计算步骤

1. 原始数据的正规化

如果有n个样品，每个样品有m个变量，其原始资料矩阵为$X=[x_{ij}]_{m\times n}$。矩阵中的每个元素x_{ij}应经过极差正规化变换为[0, 1]区间的数值，即

$$y_{ij}=\frac{x_{ij}-\min\limits_{1\leq j\leq n} x_{ij}}{\max\limits_{1\leq j\leq n} x_{ij}-\min\limits_{1\leq j\leq n} x_{ij}}$$

$$(i=1, 2, \cdots, m; \ j=1, 2, \cdots, n)$$

而得到正规化后的矩阵

$$Y=[y_{ij}]_{m\times n}$$

2. 计算变差矩阵

变差矩阵 $D=[d_{ij}]_{n\times n}$

其中 $$d_{ij}=\sum_{b=i}^{j}\sum_{a=1}^{m}(y_{ab}-\bar{y}_a(i, j))^2 \qquad (i, j=1, 2\cdots, n)$$

而 $$\bar{y}_a=\sum_{b=i}^{j}y_{ab}/(j-i+1)$$

显然有 $$d_{ij}=\begin{cases}0 & (i=j) \\ d_{ij} & (i\neq j)\end{cases}$$

所以，只要计算D的上三角部分即可以得到如下矩阵

$$D=\begin{pmatrix} d_{12} & d_{13} & \cdots & d_{1n} \\ & d_{23} & \cdots & d_{2n} \\ & & \cdots & \\ & & & d_{(n-1)n} \end{pmatrix}$$

3. 计算最优二分割

由矩阵D可以计算出所有的二分割的总变差，亦即对每一个p（$p=n, n-1, \cdots, 2$）计算相应的总变差

$$S_p(2,\ j) \qquad (j=1,\ 2,\ \cdots,\ p-1)$$

找出最小值，确定各子段的最优二分割点$t1(p)$，即

$$S_p(2,\ t1(p))=\min_{1\leq j\leq p-1} S_p(2,\ j)$$

则得到n个样品的最优二分割

$$\{x_1,\ x_2,\ \cdots,\ x_{t1(n)}\}\{x_{t1(n)+1},\ \cdots,\ x_n\}$$

4. 计算最优三分割

对于$p=n,\ n-1,\ \cdots,\ 4,\ 3$，可由$S_j(2,\ t1(j))(j=2,\ 3,\ \cdots,\ p-1)$，以及矩阵$D$分别计算

$$S_p(3,\ t1(j),\ j)=S_j(2,\ t1(j)+d_{(j+1)p}$$
$$(p=n,\ n-1,\ \cdots,\ 4,\ 3)$$

然后找出最小值，即

$$S_p(3,\ t1(p),\ t2(p)=\min_{2\leq j\leq p-1}S_p(3,\ t1(j),\ j)$$

则得到n个样品的最优三分割

$$\{x_1,\ x_2,\ \cdots,\ x_{t1(n)}\}\{x_{t1(n)+1},\ \cdots,\ x_{t2(n)}\}\{x_{t2(n)+1},\ \cdots,\ x_n\}$$

5. 计算最优k分割

可以按类似的办法，由最优三分割计算最优四分割；再由最优四分割计算最优五分割；……，如果已经得到了最优$k-1$分割，则可计算最优k分割。此时，应对$p=n,\ n-1,\ \cdots,\ k$分别计算

$$S_p(k,\ t1(j),\ \cdots,\ t(k-2)(j),\ j)$$
$$=S_j((k-1),\ t1(j),\ \cdots,\ t(k-2)(j))+d_{(j+1)p}$$
$$(p=n,\ n-1,\ \cdots,\ k;\ j=k-1,\ k-2,\ \cdots,\ p-1)$$

找出最小值，就可以确定最优k分割。

四、平面邻接样品的分割

前面所讨论的是一维有序样品的分割问题，就是对一条直线上的样品进行分段的问题。然而在地质研究工作中，经常会遇到平面（二维）上的有序样品的分割问题。在平面上谈“有序”，既不准确又不方便。而是应该将有序转换为“邻接”的概念更为合适。地质学中的沉积相带划分，地质构造上的分区，探区中的含油气地质条件的分带等等都属于平面邻接样品的分割问题。

如果有n个样品，每个样品观测了m个变量$x_{i1},\ x_{i2},\ \cdots,\ x_{im}(i=1,\ 2,\ \cdots,\ n)$，第$i$个样品在平面上的坐标为$y_{i1},\ y_{i2}$。那么，$m+2$维的向量（$y_{i1},\ y_{i2},\ x_{i1},\ x_{i2},\ \cdots,x_{im}$）就是第$i$个样品的全部信息。现在的问题是将$n$个样品分成$k$类，并使每类内部样品之间的差异最小，而类与类之间的差异最大，则认为是这种分类是平面邻接样品的最优分割。

平面上邻接样品的分割问题要比一条直线有序样品的分割问题复杂的多。目前，虽然有些人给出一些算法，但是，都不很完善，还不具备解决实际问题的能力。

五、算　　例

四川盆地中某口探井的2670～2720m井段，有6种测井资料，即：自然伽马、中子、密度、声波时差、浅侧向电阻率，深侧向电阻率，这6种测井数据可作为6个变量。在2670～2720m井段内，以上6种测井曲线均有明显的变化，可划分为56个小段，每个小段可作为一个样品。这56个样品的6个变量的采样值见表2-3-8。

表 2-3-8　六种测井资料的原始采样数据

样品号	自然伽马	中　子	密　度	声波时差	浅侧向	深侧向
1	28	3	2.84	49	700	700
2	60	7.5	2.78	53	90	65
3	10	0.75	2.92	50	1100	1000
4	48	0	2.83	52	110	88
5	10	1	2.86	50	1600	1900
6	35	9.75	2.74	48	100	150
7	16	[illegible]	2.70	48	140	155
8	25	12	2.67	49	28	28
9	52	3	2.68	51	110	100
10	24	0.75	2.72	52	200	200
11	20	0	2.71	48	1300	1300
12	25	1.5	2.75	49	1000	1000
13	45	4.5	2.83	48.5	220	240
14	3	−1.5	2.96	51.5	1800	1300
15	26	−1.6	2.89	52	1600	1650
16	45	−1.4	2.92	52	1900	1850
17	4	3	2.80	50	1000	1500
18	5	10.6	2.58	56	95	160
19	10	10	2.60	57	54	80
20	19	2.2	2.68	50	150	300
21	15	4.6	2.66	52.5	80	135
22	20	7	2.57	56	55	80
23	19	7.6	2.65	55	40	95
24	15	11	2.65	61	25	33
25	10	5.5	2.67	53.5	78	100
26	16	4.5	2.63	52.5	40	54
27	14	0.75	2.66	48.5	1000	800
28	17	1.5	2.65	48.5	1600	1400
29	15	2.3	2.67	48.5	500	650
30	25	6.7	2.70	48	160	210
31	16	1.5	2.68	47.5	400	500
32	12	3	2.69	47.5	300	550
33	17	4.5	2.70	48	210	300
34	11	1.5	2.71	42	400	560
35	13	0	2.68	42	300	500

续表

样品号	自然伽马	中子	密度	声波时差	浅侧向	深侧向
36	20	0.75	2.68	42	560	670
37	20	0.8	2.70	42	400	560
38	15	3.8	2.75	41.5	570	680
39	14	4.5	2.825	41	350	430
40	13	4.3	2.774	41	400	440
41	24	4.5	2.76	41	650	700
42	21	−0.8	2.96	50	1900	1600
43	25	1.5	2.89	54	900	700
44	10	0.75	2.90	60	1250	830
45	10	0	2.90	52.5	1000	750
46	15	8.8	2.75	55	50	32
47	79	−0.3	2.935	51.5	1500	960
48	9	11.8	2.79	60	23	23
49	45	10.5	2.775	60	30	90
50	15	4.3	2.84	54	115	30
51	64	17.8	2.64	58	30	26
52	62	11	2.70	57	54	28
53	94	18	2.725	60	26	18
54	53	13.5	2.80	57.5	34	21
55	61	12	2.775	60	29	17
56	40	9	2.835	57.5	31	24

在本算例中，对表2-3-8中的数据进行了2～20次分割，计算结果如下：

2分割：(1-47)，(48-56)

3分割：(1-41)，(42-47)，(48-56)

4分割：(1-17)，(18-26)，(27-47)，(48-56)

5分割：(1-13)，(14-17)，(18-26)，(27-47)，(48-56)

6分割：(1-13)，(14-17)，(18-26)，(27-41)，(42-47)，(48-56)

7分割：(1-13)，(14-17)，(18-26)，(27-28)，(29-41)，(42-47)，(48-56)

8分割：(1-4)，(5-5)，(6-13)，(14-17)，(18-26)，(27-41)，(42-47)，(48-56)

9分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-17)，(18-26)，(27-41)，(42-47) (48-56)

10分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-17)，(18-26)，(27-28)，(29-41)，(42-47)，(48-56)

11分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-17)，(18-26)，(27-41)，(42-45)，(46-46)，(47-47)，(48-56)

12分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-17)，(18-26)，(27-28)，(29-41)，(42-45)，(46-46)，(47-47)，(48-56)

13分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-17)，(18-26)，(27-28)，(29-41)，(42-45)，(46-46)，(47-47)，(48-50)，(51-56)

14分割：(1-4)，(5-5)，(6-10)，(11-12)，(13-13)，(14-17)，(18-26)，(27-28)，(29-41)，(42-45)，(46-46)，(47-47)，(48-50)，(51-56)

15分割：(1-4)，(5-5)，(6-10)，(11-12)，(13-13)，(14-17)，(18-26)，(27-28)，(29-33)，(34-41)，(42-45)，(46-46)，(47-47)，(48-50)，(51-56)

16分割：(1-4)，(5-5)，(6-10)，(11-12)，(13-13)，(14-17)，(18-26)，(27-28)，(29-33)，(34-41)，(42-42)，(43-45)，(46-46)，(47-47)，(48-50)，(51-56)

17分割：(1-4)，(5-5)，(6-10)，(11-12)，(13-13)，(14-16)，(17-17)，(18-26)，(27-28)，(29-33)，(34-41)，(42-42)，(43-45)，(46-46)，(47-47)，(48-50)，(51-56)

18分割：(1-4)，(5-5)，(6-10)，(11-12)，(13-13)，(14-16)，(17-17)，(18-26)，(27-28)，(29-33)，(34-41)，(42-42)，(43-45)，(46-46)，(47-47)，(48-50)，(51-53)，(54-56)

19分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-16)，(17-17)，(18-26)，(27-28)，(29-33)，(34-41)，(42-42)，(43-45)，(46-46)，(47-47)，(48-49)，(50-50)，(51-52)，(53-53)，(54-56)

20分割：(1-4)，(5-5)，(6-10)，(11-13)，(14-16)，(17-17)，(18-26)，(27-28)，(29-33)，(34-41)，(42-42)，(43-45)，(46-46)，(47-47)，(48-49)，(50-50)，(51-51)，(52-52)，(53-53)，(54-56)

从以上计算结果看出，5次分割时，已将大段不同岩性分开，分割点分别为13、17、26、47号样品，在以上20次分割计算中这些分割点均使用了16次以上，说明这些分割点的确是重要的岩性分界点。采用17次分割时，已可以划分出不同的岩性段。到20分割时，可将岩性划分的更为详细。见图2-3-6。

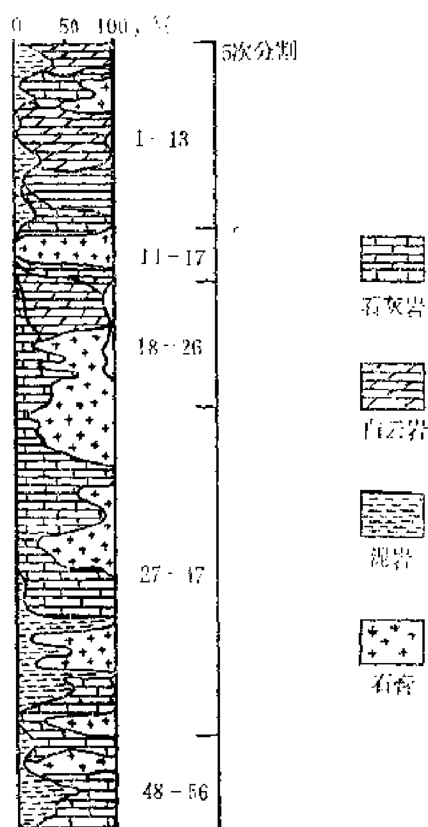


图 2-3-6 测井解释剖面与一维有序分割计算结果对比图

第四节 动态聚类分析

动态聚类是首先粗糙地进行初始分类，然后再逐步调整，直至分类结果满意为止，所以，动态聚类法有时也称作逐步聚类。但是，这里的“逐步”的含义并不是逐步筛选变量，而是逐步调整分类。

一、动态聚类的过程

动态聚类的过程见图2-3-7。为了进行粗糙的初始分类，首先要选一批有代表性的样品作为“凝聚点”，然后让其他样品按某种原则向凝聚点汇聚，便可得到一个初始分类。

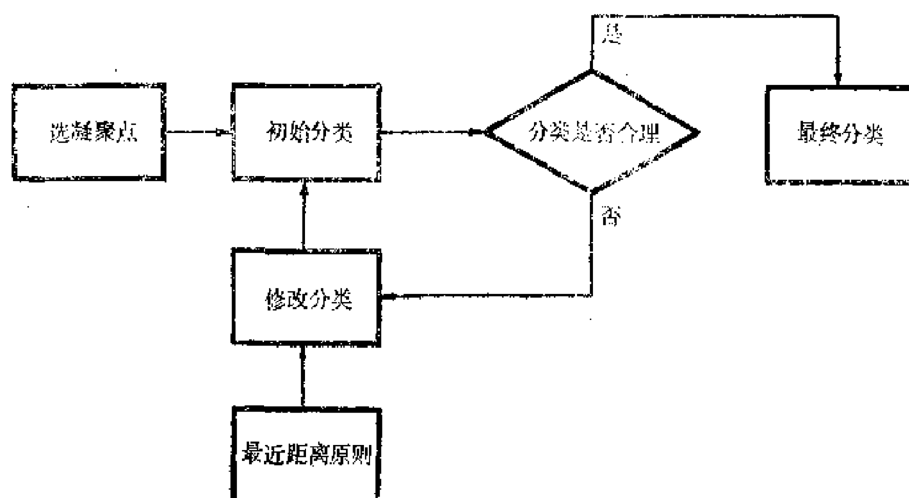


图 2-3-7 动态聚类的过程

有了初始分类之后，就要判断初始分类是否合理，如果不合理就修改原分类，如果修改后仍然不合理，就再一次修改分类，一直到分类合理为止。

动态聚类的计算工作量较小，占用计算机内存也较少，方法简单易行，故当样品数量较大时，采用动态聚类法是比较合适的。由于图2-3-7中的每个环节都可以有不同的处理方法，因而也就产生了种类繁多的动态聚类方法。

二、凝聚点与初始分类

1. 凝聚点

凝聚点就是一批有代表性的样品点（或变量点），这些点被当成将要形成类的中心，选择凝聚点的方法很多，通常采用如下办法：

（1）凭经验选择凝聚点。对于有经验的地质人员来说，得到一批数据后，大体上如何分类，分成几类，一般都会有个初步想法。并且，凭经验也可以从每一类中挑选出一个有代表性的样品点作为该类的凝聚点。或者凭经验将全部数据分成k个类，计算每个类的重心，并将这些重心作为各类的凝聚点。

（2）用密度法选择凝聚点。如果以每个样品作为球心，以某个正数d为半径作球，落在这个球内的样品数称为该样品点的密度。或者事先确定一个自然数k，以某一个样品为球心作球，如果作成一个恰好包含k个样品点在内的球，这个球的半径的倒数可称为该样品点的密度。

当所有样品点的密度都算好后，首先选择密度最大的样品点作为第一个凝聚点。并且人为地确定一个正数D，然后选次大密度的样品点，若它与第一个凝聚点的距离大于D，就将它作为第二个凝聚点；若距离小于D，就不能作为凝聚点。这样，按密度大小依次进行考

察，凡与已选中的各凝聚点之间的距离均大于D的则作为凝聚点，否则就不能作为凝聚点。

（3）用均值法选择凝聚点。预先人为地确定一个正数D，首先选择全部样品的均值作为第一个凝聚点。然后将样品依次输入，如果输入的样品与已经确定的各个凝聚点之间的距离都大于D，就可以选择该样品点作为新的凝聚点，否则就不能作为凝聚点。

（4）按编号选取凝聚点。确定一个正整数k，然后，按样品的编号取前k个样品点作为凝聚点。

有了凝聚点之后，形成初始分类也有不同的方法；并且，形成初始分类也不一定非通过凝聚点不可。下面介绍的就是一些常用的形成初始分类的方法。

（1）选择一批凝聚点后，其他样品向距离最近的凝聚点归类，这是最常用的形成初始分类的方法。

（2）选择一批凝聚点后，将样品依次归入与其距离最近的凝聚点那一类，然后重新计算该类的重心，并且用重心代替原凝聚点，再计算下一个样品的归类，直到所有的样品都归到相应的类中为止。

（3）先人为地规定一个正数d，选择$G_1=\{x_1\}$，计算样品x_2与x_1之间的距离d_{21}，如果$d_{21}<d$，则将x_2归入G_1类，否则就建立新类$G_2=\{x_2\}$。当某一步轮到输入x_l时，假如当时已形成了k个类G_1，G_2，…，G_k，每个类第一次进去的样品记作x_{i1}，x_{i2}，…，x_{ik}（显然$i1=1$），如果$d_{l,ij}>d$（$j=1$，2，…，k），则将x_l建立为新的第$k+1$类，即$G_{k+1}=\{x_l\}$；否则就将x_l归入与x_{i1}，x_{i2}，…，x_{ik}距离最近的那一类。

这种方法的优点是计算速度快，每个样品只需要通过一次计算，其缺点是分类结果与样品的排列次序有关，而且先建立类容易比后建立的类收容更多的样品。

（4）用任何一种聚类方法所得的结果作为初始分类。

三、按批修改法

一般情况下，初始分类都是不合理的分类，因此需要进行调整，调整的方法很多，按批修改法是其中的一种，其调整步骤如下：

（1）选择一批凝聚点并且确定分类统计量；

（2）将所有的样品按与其距离最近的凝聚点归类；

（3）计算每一类的重心，将重心作为新的凝聚点。如果计算后新产生的凝聚点与前一次的原有凝聚点重合，则分类结束；否则，回到步骤2，直到与前次的凝聚点重合为止。

按批修改法的计算过程中，每一步修改都会使选定的分类函数值逐渐缩小，最后趋于定值，亦即这个计算过程是收敛的，多数情况下收敛速度较快。有时为了节约计算时间，也可以规定迭代次数不超过L次，例如$L=5\sim10$。

按批修改法的思路很象计算方法中的迭代法、不过迭代法不受初始值的影响，而按批修改法却不然，不同的初始分类可能得到不同的结果。这是动态聚类法的缺点。

四、逐个修改法

按批修改法是等样品全部调整完毕后才改变凝聚点，而逐个修改法是每个样品一旦调整后立即改变凝聚点。逐个修改法的计算步骤如下：

（1）人为地确定分类数k，取前k个样品作为凝聚点；

（2）将剩下的（$n-k$）个样品逐个归入与其距离最近的凝聚点那一类，随即计算该类的重心，并用这个重心代替原凝聚点；

（3）将n个样品重新按步骤2逐个归类。如果n个样品此时所属的类与原来的归类完全一样，则分类结束，否则重复步骤3。

五、其他种动态聚类方法

在按批修改法和逐个修改法基础上，产生了动态聚类的许多变种。这些方法可以使在调整类的过程中，相距很近的类可以合并，而包含样品数目过多的大类也可以分裂。也就是说，不仅每个样品的归属要不断调整，并且类的数目随着分类过程的发展也在变化。

1．三参数法

（1）人为地选定三个控制参数k、C、R。

（2）把前k个样品选作凝聚点，使它们两两之间的距离均大于或等于C。为此，要计算这k个凝聚点之间的距离，如果最小的距离小于C，则将这两个凝聚点合并，并用这两个点的重心作为新的凝聚点。重复这个步骤，直至所有凝聚点之间的距离均大于或等于C为止。

（3）对剩下的（$n-k$）个样品逐个地计算其与所有凝聚点的距离，如果其中最小的距离大于R，则将该样品点作为新的凝聚点，自成一个新类；如果最小距离小于或等于R，则将该样品归入与其距离最近的凝聚点的那一类，随即重新计算这一类的重心，并以此重心作为新的凝聚点。

重新检查凝聚点之间的距离，如果有小于C的，用步骤（2）的办法将相应的两类合并，直到所有凝聚点之间的距离均大于或等于C为止。

（4）对n个样品用步骤（3）的原则逐个地处理。如果新的分类与上一次完全相同，则聚类过程结束，否则重复步骤（1）。

2．四参数法

（1）选择四个控制参数T、MZ、MT、MC。

（2）进行初始分类，计算各类的重心。

（3）设k为分类过程中类的数目，由于有合并或分裂的缘故，k与初始分类数目不一定相同。对每一个样品依次计算其与各类重心的距离。这里有三种情况：

①如果样品至各类（包括样品所属类）重心的最小距离大于T，则将该样品放入未分类的剩余样品集合中，计算失去该样品原属类的重心；

②如果样品至某些类的重心的最小距离小于或等于T，但样品并不属于与其距离最近的那一类，则将该样品划入与其距离最近的那一类，并分别计算失去和获得该样品的那两类的重心；

③如果样品在剩余集合中，且至某些类重心的最短距离小于或等于T，则将该样品划入与其距离最近的重心所属的那一类，并且重新计算样品所归入类的重心。

（4）当对所有样品执行完步骤（3）以后，检验各类的样品数，把样品数目小于MZ的类全部划入剩余集合。

（5）重复步骤（3）与（4）MT次。若不足MT次分类即已收敛，则转入步骤（6）；

若达到MT次仍不收敛，也要转入步骤（6）。

（6）将最近的两类合并，重复步骤（2）~（5），直到类的数目达到MC为止。

3. 七参数法

（1）选择七个控制参数NS、ND、NT、TN、TE、TC、和IX。

（2）选择一批凝聚点。

（3）将样品按成批修改法归类，并规定成批修改法的次数不超过NS次，若不足NS次分类即已收敛则转入步骤（4），否则要在作完NS次之后转入步骤（4），而不必达到收敛。

（4）如果某一类的样品数小于TN（例如可令$TN=2$），则将该类取消，并且该类样品不参加下面的运算。

（5）按以下原则将类进行合并或分裂：

①如果类的数目k大于或等于$2ND$，则进入合并的迭代过程；

②如果类的数目k小于$0.5ND$，则进入分裂的迭代过程。

（6）重新计算各类的重心，作为新的凝聚点，将全部样品按步骤（3）所述的方法进行聚类。

（7）重复步骤（4）、（5）、（6），最多达IX次，或至过程收敛。

在分裂的迭代过程中，在某一类G_i中，如果某个变量x_j的标准差$\sigma_j^{(i)}>\sigma_j \cdot TE$，其中$\sigma_j$为全部样品的变量$x_j$的标准差，则将$G_i$分裂为两类，并以$x_j$在$G_i$中的均值来划线，均值以上的属一类，其余的为另一类。分别计算两个类的重心，这两个重心之间的距离如果大于$1.1TC$，则可以进行分裂，否则不作分裂。

在合并的迭代过程中，计算各类重心之间距离时，若两个最近重心之间的距离小于TC，则合并这两类，并计算这个新类与其他各类重心之间的距离，这一过程在每次迭代过程中最多进行NT次。显然，如果不具备合并的条件，当然不进行合并，因而不是每次迭代都发生NT次合并。

第五节 聚类预报

聚类分析主要用于分类，但也可以利用分类结果进行预报。目前，主要是用回归分析和判别分析进行预报，但是，由于聚类预报有其独特之处，可以弥补回归预报和判别预报的不足，因此，聚类预报是值得重视的一种方法。

一、AID（Automatic Interaction Detection）法

回归分析是最常用的定量预报方法。但回归预报经常遇到一些困难，例如，有多个自变量的回归方程，当实际规律不是线性关系时，由于寻找非线性关系比较困难，经常不得不用线性模型或多项式模型来代替。这样，由于模型不准确必然会增加预报的误差。另外，回归预报是一种基于平滑平均思想的预报，因而，对于某些特殊情况，特别是突变事件，例如，大地震、台风、特富矿段等等往往预报不准。而这些特殊情况往往是人们所最关注的，故由于预报不准会造成很大的损失。而聚类预报恰好可以弥补回归预报的这些不足。

为了说明聚类预报的基本思想，这里先看一个简单例子。自变量x与因变量y共有11对

观测值，如下：

$x=1$	2	3	4	5	6	7	8	9	10	11
$y=1.4$	2.0	2.7	3.1	13.5	16	14	4.9	5.6	6.2	6.6

我们通过这批数据来建立聚类预报关系。

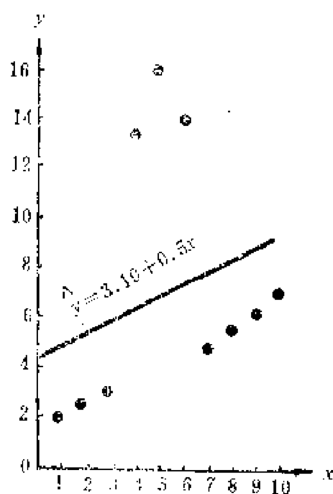


图 2-3-8　回归预报方程

根据这批数据，可作自变量x与因变量y之间的关系图，见图2-3-8。可以发现，前面的4个点与后面的4个点大致在一元线性回归方程$\hat{y}_0=1+0.5x$附近，而中间3个点显著偏高。如果用回归分析方法，根据11个数据点建立预报方程，其一元线性回归方程为

$$\hat{y}=3.91+0.5x$$

由图2-3-8可看出，用回归方法进行预报效果很不好。对于这个例子采用多项式回归预报，效果也不好。

解决这个问题的一个很朴素的想法就是把x分成几个区间，将对应于这个区间的y值进行平均，这个平均值就当作这个区间的预报值。对于新给定的自变量x值，它落在哪个区间，就以这个区间的平均值$\bar{y}$，作为因变量y的预报值。

对于前面的数据，可将x分成3个区间，［0.5，4.5］、［4.5，7.5］、［7.5，11.5］，分别计算落入这些区间的y_i（$i=1,2,3$）的平均值：

$$\bar{y}_1=\frac{1}{4}(1.4+2.0+2.7+3.1)=2.3$$

$$\bar{y}_2=\frac{1}{3}(13.5+16+14)=14.5$$

$$\bar{y}_3=\frac{1}{4}(4.9+5.6+6.2+6.6)=5.825$$

如果现在有一个新的x观测值，例如$x=3$，显然，它落在第一区间，因而，就可以用$\bar{y}_1=2.3$来预报与它对应的y值。

需要指出，自变量x的区间划分必须根据y值的大小来确定。如果把$\{y_1, y_2, \cdots, y_{11}\}$看成为一个一维有序样品，利用最优分割法可以将它们分成k段，这k段对应的x正好是k个区间。

此处采用一维有序分割的简单算法，即用每次二分段法进行如下计算：

$$\bar{y}=\frac{1}{11}\sum_{i=1}^{11}y_i=6.91$$

将前n_1（$1\leqslant n_1\leqslant 10$）个样品归为一类，计算它们的均值$\bar{y}_1$，以及分类函数$E$

$$E=\frac{n_1 n}{n_2}(\bar{y}_1-\bar{y})^2$$

式中的$n_2=n-n_1$，$n=11$。计算的目的是找出E的极大值，计算结果见表2-3-9。

由表2-3-9中可见，E的极大值为133.53，这表明从y_4处分段，即$\{y_1, y_2, y_3, y_4\}$，$\{y_5, y_6, y_7, y_8, y_9, y_{10}, y_{11}\}$为好。

表 2-3-9 第一次分段的E值

n_1	1	2	3	4	5	6	7	8	9	10
$\bar{y}_1$	1.40	1.70	2.03	2.30	4.54	6.45	7.53	7.20	7.02	6.94
E	33.38	66.33	98.19	133.53	51.44	2.78	7.43	2.48	0.61	0.11

然后将$\{y_1\sim y_4\}$，$\{y_5\sim y_{11}\}$分别试分为两类，看哪段的E值更大。这两段E的最大值分别为1.44与129.66，后者大，说明应将第二段再分为两段，按同样方法计算，应分成$\{y_5, y_6, y_7\}$，$\{y_8, y_9, y_{10}, y_{11}\}$。然后再将已得到的三类各自试分成两类，再划分$E$值最大的那一类。可以按这个办法继续分割下去。结束这个过程有两种控制方法，一种是规定类的个数，达到规定的类数就结束；另一种计算方法是求

$$ESP_i=\frac{SSQ_i}{SSQ}$$

式中的SSQ_i为第i类的离差平方和，SSQ为总的离差平方和。当$ESP_i<ESP$（事先给定的阈值）对一切的i均成立时，过程结束。

现在采用第二个办法，取$ESP=1.5\%$，经计算类的离差平方和分别为

$$SSQ_1=1.7$$
$$SSQ_2=3.5$$
$$SSQ_3=1.6475$$

总的离差平方和$SSQ=269.4$。而

$$ESP_1=0.63\%$$
$$ESP_2=1.30\%$$
$$ESP_3=0.61\%$$

均小于1.5，所以过程结束。如果取$ESP=1\%$，则第二类还必须再进行分段。

如此，将y分成三类，即$\{y_1, y_2, y_3, y_4\}$，$\{y_5, y_6, y_7\}$，$\{y_8, y_9, y_{10}, y_{11}\}$，这三段对应的$x$分段为$\{x_1, x_2, x_3, x_4\}$，$\{x_5, x_6, x_7\}$，$\{x_8, x_9, x_{10}, x_{11}\}$。为使$x$的分段之间是彼此邻接的，第一类与第二类的分界点可取成两个端点的平均数，即$(4+5)/2=4.5$；类似的，第二类与第三类的分界点为$(7+8)/2=7.5$。从而得到x所分成的三个区间

$$\{x\leqslant4.5\}, \{4.5<x\leqslant7.5\}, \{x>7.5\}$$

这三个区间的y的平均值为

$$\bar{y}_1=2.3, \bar{y}_2=14.5, \bar{y}_3=5.825$$

这种聚类预报方法很简单，但对某些问题来说，预报效果比回归预报更好一些。

当影响y的自变量不止一个，而有m个时，也能用上述的聚类方法进行预报，这样方法通常称为AID法。自变量为$x_1, x_2, \cdots, x_m$时，是设法在其变化范围内，分成许多长方体，计算落在这些长方体内的y的平均值，对于新给定的x值，落在哪个长方体内，就用该长方体内的y的平均值来预报。AID法给出了划分这些长方体的一个有效办法，其计算步骤是：

（1）将数据按第j（$1\leqslant j\leqslant m$）个自变量的次序由小到大重新排列，使$\{y_i\}$得到一个新的次序，并看作是这个次序上的一维有序样品。将样品作最优二分段，相应的分类函数E的

极大值记作$SSQ_i(t_i)$，其中t_i为达到极大值的下标（按新的次序）。在$SSQ_1(t_1)$，$SSQ_2(t_2)$，…，$SSQ_m(t_m)$中取极大，比如$SSQ_{i_1}(t_{i_1})$达到极大，则第一次就用变量x_{i_1}，将自变量的变化范围分成两块

$$G_1=\left\{\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}\middle|\ x_{i_1}\leqslant z_1\right\},\quad G_2=\left\{\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}\middle|\ x_{i_1}>z_1\right\}$$

其中
$$z_1=\frac{x_{t_{j_1}i_1}+x_{t_{i_1+1}i_1}}{2}$$

z_1就是第j_1个自变量在类G_1和类G_2中最接近的两点的均值。

（2）对类G_1和类G_2各自进行最优二分段，对分类函数较大者，（设为G_2）先分割。分割G_2用的是变量x_{i_2}（j_2仍按上面的方法确定）分割下标为t_{i_2}，分类函数为$SSQ_{i_2}(t_{i_2})$，分割点为

$$z_2=\frac{x_{t_{j_2}j_2}x_{t_{j_2+1}j_2}}{2}$$

分成的两类分别记作G_3和G_4。即

$$G_3=\left\{\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}\middle|\ x_{i_2}\leqslant z_2\right\},\quad G_4=\left\{\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}\middle|\ x_{i_2}>z_2\right\}$$

然后对G_1、G_3、G_4重复步骤2。可用前面介绍的两种办法结束该过程。

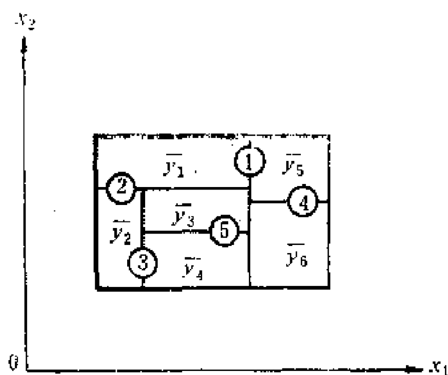


图 2-3-9　AID法示意图

为了便于理解AID法，图2-3-9给出了$m=2$的一个示意图。①，②，…，⑤表示第一，二，…，五次分割的界限。

由图2-3-9可知，第一次分割用x_1，第二次用x_2，第三次用x_1，第四次用x_2，第五次还是x_2。经五次分割将自变量变化范围分成六块，每一块可以确定预报量$\hat{y}_1$，$\hat{y}_2$，…，$\hat{y}_6$，从而就可以对样品预报。

AID法每次分割只用一个变量，使得计算十分方便，而且还具有筛选变量的功能，这些都是AID法的优点，也是使AID法得到广泛采用的原因。

二、GAID法

AID法每次分割只用一个变量，当单个变量与y的相关性不大，而几个变量与y的复相关很大时，AID法的计算结果有时很不理想。方开泰等人（1979）对AID法作了改进，使得每一次对自变量区域的分割可以依赖于多个变量，并把改进后的方法称作GAID法。对于多个自变量，AID法的分割结果，自变量的区域都是超多面体，而多面体的每个面都平行于某个

坐标轴。而GAID法分割的自变量区域也是多面体，但是，每个面不一定平行于任何坐标轴。图2-3-10给出了两个变量时，即$m=2$的分割示意图。第一、二次分割都用了两个变量（不平行于任何坐标轴），第三次分割是用变量x_1（平行于坐标轴x_2），第四次分割是用变量x_2（平行于坐标轴x_1）。最后将自变量区域分成为五块，计算每一块上的$\{y_i\}$均值，得到$\bar{y}_1$，$\bar{y}_2$，$\bar{y}_3$，$\bar{y}_4$，$\bar{y}_5$。新的样品落到区域上的哪一块，就以该块的均值作为它的预报值。

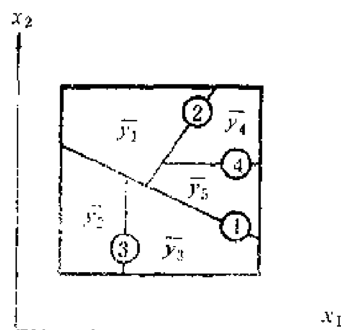


图 2-3-10　GAID法示意图

如果G中有n个样品，每个样品有m个变量。GAID法是在某个分类函数之下将G分裂为G_1、G_2两类，分别计算G_1、G_2的重心$\bar{x}_1$、$\bar{x}_2$。将n个样品x_1，x_2，…，x_n向由$\bar{x}_1$、$\bar{x}_2$所决定的直线上投影，按投影值的大小将样品重新编号，得到$x_{(1)}$，$x_{(2)}$，…，$x_{(n)}$。

GAID法的具体计算步骤如下：

（1）将数据按第j（$1\leqslant j\leqslant m$）个自变量的次序重排，按这个次序把$\{Y_i\}$看成有序样品，将它最优二分段，相应的分类函数记作SSQ_j。若j_1使

$$u_1=\max_{1\leqslant j\leqslant m} SSQ_j$$

相应的分割下标为t_{j_1}。令$p_{(1)}\leqslant p_{(2)}\leqslant\cdots\leqslant p_{(n)}$为$x_{1_{j_1}}$，$x_{2_{j_1}}$，…，$x_{n_{j_1}}$按由小到大重排后的数值。

（2）如已选入了r个变量x_{i_1}，x_{i_2}，…，x_{i_r}，再加上未选中的一个变量x_i（$i=1$，2，…，m；$i\neq j_1$，j_2，…，j_r），使n个样品向$r+1$变量所决定的方向投影，并按投影值将样品次序重排，再以这个次序将$\{Y_i\}$最优二分割。相应的分类函数记作$SSQ_{i_1,\cdots,i_r,i}$，令

$$u_{r+1}=\max_{\substack{i=1,2,\cdots,m\\ i\neq i_1,\cdots,i_r}} SSQ_{i_1,\cdots,i_r,i}$$

并设$i=j_{r+1}$时达到极大值，相应的分割下标为。$t_{i_{r+1}}$。

如果$u_{r+1}\leqslant u_r$，表示新增变量后并未使分类函数有任何增加，就以原选进的r个变量x_{i_1}，…，x_{i_r}所决定的可分方向v来作分割。n个样品（用选中的r个变量）在v上的投影按其大小排列记为$t_{(1)}\leqslant t_{(2)}\leqslant\cdots\leqslant t_{(n)}$，$\{Y_i\}$按这个次序最优二分割的下标为$t_r$，则所分成的两类是：

$$G_1=\left\{\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}\middle|(x_{i_1},\ \cdots,\ x_{i_r})v\leqslant z,\right\}$$

$$G_2=\left\{\begin{pmatrix}x_1\\x_2\\\vdots\\x_m\end{pmatrix}\middle|(x_{i_1},\ \cdots,\ x_{i_r})v\leqslant z,\right\}$$

其中$z_r=\frac{1}{2}(p_{(i_r)}+p_{(i_{r+1})})$，并转入下个步骤3。

如果$u_{r+1}>u_r$，说明增加变量后有好处，再回到步骤2看能否再增加新的变量。

（3）分别计算G_1与G_2的离差平方和S_1与S_2以及G的离差平方和S，如果$(S_i/S)\leqslant ESP$（$i=1$，2）（ESP为给定的阈值），则表明它们已无须再分割，并转到步骤4。如果对某些i有$(S_i/S)>ESP$，则对凡是超过ESP的类再回到步骤1。

（4）如此自变量分为k类：G_1，G_2，…，G_k，计算落入相应各区域$\{Y_i\}$的均值$\overline{Y}_1$，$\overline{Y}_2$，…，$\overline{Y}_k$。新的样品落入哪个区域就以相应的均值来预报。

第四章　判别分析

在地质研究工作中，经常碰到样品的类型归属问题。例如，一个地质剖面中的某个层段是油层，还是水层；某块岩石标本是属于上第三系还是属于下第三系；探区中地质圈闭的含油气地质条件是有利的，还是不利的，等等。这些问题都是样品类型的归属问题。

判别分析是一种常用的地质多元统计方法，其原理是根据一组已经确知归属类型的样品，建立样品的归属类型与地质变量间的定量关系，即建立判别方程。该判别方程可以给出类型归属的界线值。对于一个新样品，可将该样品的地质变量观测值代入判别方程，求得该样品的判别值，再把这一判别值与类型归属的界线值进行比较，最后确定这个新样品的类型归属。这就是两组判别分析的基本原理。

从算法上，判别分析通常分为两组判别分析、多组判别分析以及逐步判别分析。

第一节　两组判别分析

两组判别分析是指样品的归属类型只有两种。例如，在一个探区已经发现一批地质圈闭，其中少数地质圈闭经过钻探证实，有的含油，有的不含油。现在的问题是，对于那些尚未钻探的圈闭，能否在钻探前判断一下每个圈闭的含油性呢？也就是说要判断一下它们属含油和不含油这两种类型中为那一种。

地质家十分清楚，判断这种问题用某个单项指标是比较困难的，故一般都是用多项指标进行综合研究作出判断。这好比人们观察事物，从多角度多侧面进行观测得出的结论，总比从一个角度、一个侧面观测得出的结论可靠。就上面所说的地质圈闭是含油或不含油的类型归属问题，如果只考虑圈闭的几何特征，则判断结论容易失误；若不仅考虑了圈闭的几何特征，还考虑了生油条件、储油条件、盖层条件等诸多因素，则得出的判断结论就不容易失误。

为了叙述上的方便，可以把含油圈闭的统计总体称作A，不含油圈闭的统计总体称作B。假如有m个地质变量x_1，x_2，…，x_m可以作为判断圈闭含油或不含油的指标，一般情况下，用其中的某一个地质变量x_i进行判别时，含油总体A与不含油总体B之间总有重叠部分。见图2-4-1。

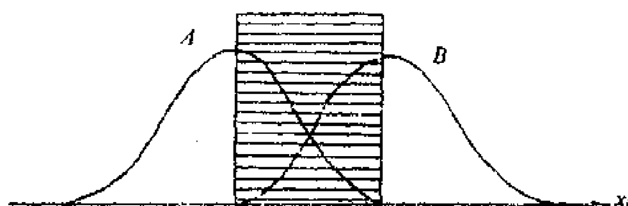


图2-4-1　用单一指标判别时总体间有重叠

由图2-4-1可见，用单一指标识别重叠部分的样品（地质圈闭）的含油性是比较困难的。如果用m个地质变量中的两个变量x_i，x_j进行判别时，则情况会有所改观，图2-4-2中的含油总体A与不含油总体B之间已不存在重叠部分，如果有，也仅仅是个别样品。图中的两个椭圆分别是A类样品（含油地质圈闭）与B类样品（不含油地质圈闭）的等

密度点轨迹。如果将两个指标x_i、x_j分别投影到直线v上，则直线v上A类样品与B类样品的重叠部分已显著减少。因而，直线v可看作是两个地质变量x_i，x_j的综合指标，或者称作线性判别函数，即

$$y=c_i x_i+c_j x_j \tag{2-4-1}$$

上式相当于三维空间（x_1，x_2，y）中通过坐标原点0的一个平面U,即,$y=c_1x_1+c_2x_2$，见图2-4-3。

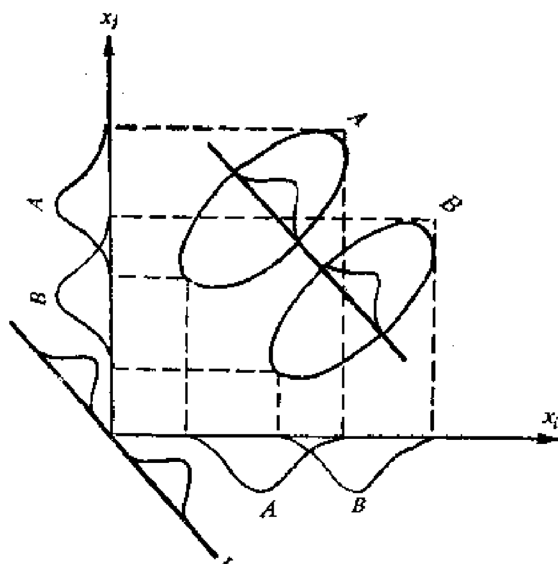


图2-4-2　用两个指标判别时两个总体间已不重叠

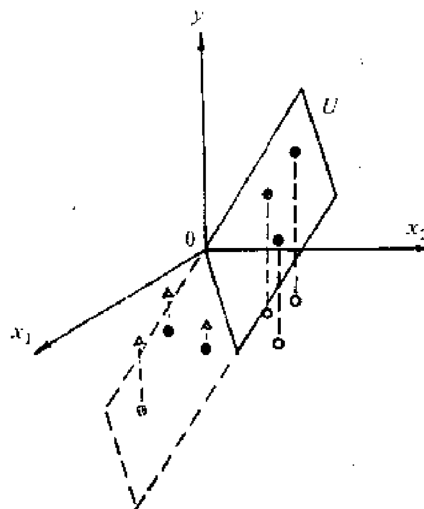


图2-4-3　A类与B类样品在平面上的投影

图2-4-3中的c_1、c_2为平面U的两个方向系数，当c_1、c_2分别取某一确定值后，平面U的空间位置则已经确定。如果c_1、c_2值选择的合适，则平面可将A类样品和B类样品区分开来。如果A类样品和B类样品分别用○和△表示，就是使A类样品点沿y轴向平面U上的投影尽量地在x_1x_2平面以上，使B类样品点沿y轴向平面U上的投影尽量地在x_1x_2平面以下，这样就能把A类样品和B类样品明显区别开来。由图2-4-3显然可见，平面U的倾角越大，则两类样品越容易区分。当平面U与x_1x_2平面越接近垂直时，两类样品的y值差距越大，但同类样品点之间的距离也被拉大。因此，我们希望选择这样的c_1、c_2值，使得平面U能够使A类样品与B类样品间的区别最大，而A类或B类内部样品间的离散性最小。这种满足两类之间最大的分离原则，在统计学上称为费歇（Fisher）准则。

一、两组线性判别方程

两组判别是指有n个样品，分别归属于第1类与第2类，每个样品有m个可供判别分类的变量，其线性判别方程的表达式为

$$y=c_1x_1+c_2x_2+\cdots+c_mx_m \tag{2-4-2}$$

式中的c_1，c_2，…，c_m为判别方程的待定系数。为确定这些待定系数，应采用使得两类样品之间的差别最大、其内部样品之间的差别最小的原则，而待定系数的值则用n个样品的观测值来确定。

这里约定用（1）表示第1类，用（2）表示第二类。对于m个变量，（1）类与（2）类分别有n_1与n_2个观测数据。即（1）类的观测数据为

$$x_{11}(1),\ x_{21}(1),\ \cdots,\ x_{m1}(1)$$
$$x_{21}(1),\ x_{22}(1),\ \cdots,\ x_{m2}(1)$$
$$\cdots\ \ \cdots\ \ \cdots\ \ \cdots\ \ \cdots\ \ \cdots$$
$$x_{1n_1}(1),\ x_{2n_1}(1),\ \cdots,\ x_{mn_1}(1)$$

而（2）类观测数据为

$$x_{11}(2),\ x_{21}(2),\ \cdots,\ x_{m1}(2)$$
$$x_{21}(2),\ x_{22}(2),\ \cdots,\ x_{m2}(2)$$
$$\cdots\ \ \cdots\ \ \cdots\ \ \cdots\ \ \cdots\ \ \cdots$$
$$x_{1n_2}(2),\ x_{2n_2}(2),\ \cdots,\ x_{mn_2}(2)$$

为了区别（1）类与（2）类，可用$\overline{y}(1)$与$\overline{y}(2)$作为分类指标。为使这两类区别最大，则要求

$$Q=[\overline{y}(1)-\overline{y}(2)]^2 \quad 为最大$$

而（1）类或（2）类内部样品之间的差别，可用下面的两式表达

$$\sum_{k=1}^{n_1}[y_k(1)-\overline{y}(1)]^2;\quad \sum_{k=1}^{n_2}[y_k(2)-\overline{y}(2)]^2$$

为使各类内部样品间的差别最小，则要求

$$F=\sum_{k=1}^{n_1}[y_k(1)-\overline{y}(1)]^2+\sum_{k=1}^{n_2}[y_k(2)-\overline{y}(2)]^2 \quad 为最小$$

为同时满足这两项要求，可用下面的表达式E作为分类的综合指标：

$$E=\frac{[\overline{y}(1)-\overline{y}(2)]^2}{\sum\limits_{k=1}^{n_1}[y_k(1)-\overline{y}(1)]^2+\sum\limits_{k=1}^{n_2}[y_k(2)-\overline{y}(2)]^2}=\frac{U}{V} \tag{2-4-3}$$

为使E值达到最大，可使E对c_1，c_2，…，c_m的偏导数等于0，而得到如下方程组

$$\begin{cases}\dfrac{\partial E}{\partial c_1}=0\\ \dfrac{\partial E}{\partial c_2}=0\\ \cdots\\ \dfrac{\partial E}{\partial c_m}=0\end{cases}$$

为计算待定系数c_1，c_2，…，c_m的值，可先对（2-4-3）式两边取对数，即

$$\ln E=\ln U-\ln V$$

对两边求c_i的偏导数，令其为0，则有

$$\frac{\partial \ln E}{\partial c_i}=\frac{\partial \ln U}{\partial c_i}-\frac{\partial \ln V}{\partial c_i}=0 \qquad (i=1,\ 2,\ \cdots,\ m)$$

$$\frac{1}{U}\frac{\partial U}{\partial c_i}=\frac{1}{V}\frac{\partial V}{\partial c_i}$$

$$\frac{1}{E}\frac{\partial U}{\partial c_i}=\frac{\partial V}{\partial c_i}$$

因为

$$\begin{aligned}U&=[\overline{y}(1)-\overline{y}(2)]^2\\&=\Big[\sum_{i=1}^{m}c_i\bar{x}_i(1)-\sum_{i=1}^{m}c_i\bar{x}_i(2)\Big]^2\\&=\Big[\sum_{i=1}^{m}c_i(\bar{x}_i(1)-\bar{x}_i(2))\Big]^2\\&=\Big[\sum_{i=1}^{m}c_id_i\Big]^2\end{aligned}\tag{2-4-4}$$

(2-4-4)式中

$$d_i=[\bar{x}_i(1)-\bar{x}_i(2)]$$

因为

$$\begin{aligned}V&=\sum_{k=1}^{n_1}[y_k(1)-\overline{y}(1)]^2+\sum_{k=1}^{n_2}[y_k(2)-\overline{y}(2)]^2\\&=\sum_{k=1}^{n_1}\Big[\sum_{i=1}^{m}c_i(x_{ik}(1)-\bar{x}_i(1))\Big]^2+\sum_{k=1}^{n_2}\Big[\sum_{i=1}^{m}c_i(x_{ik}(2)-\bar{x}_i(2))\Big]^2\\&=\sum_{k=1}^{n_1}\sum_{i=1}^{m}c_i[x_{ik}(1)-\bar{x}_i(1)]\cdot\sum_{j=1}^{m}c_j[x_{jk}(1)-\bar{x}_j(1)]\\&\quad+\sum_{k=1}^{n_2}\sum_{i=1}^{m}c_i[x_{ik}(2)-\bar{x}_i(2)]\cdot\sum_{j=1}^{m}c_j[x_{jk}(2)-\bar{x}_j(2)]\\&=\sum_{i=1}^{m}\sum_{j=1}^{m}c_ic_j\Big[\sum_{k=1}^{n_1}(x_{ik}(1)-\bar{x}_i(1))(x_{jk}(1)-\bar{x}_j(1))\\&\quad+\sum_{k=1}^{n_2}(x_{ik}(2)-\bar{x}_i(2))(x_{jk}(2)-\bar{x}_j(2))\Big]\\&=\sum_{i=1}^{m}\sum_{j=1}^{m}c_ic_js_{ij}\end{aligned}\tag{2-4-5}$$

(2-4-5)式中

$$\begin{aligned}s_{ij}=&\sum_{k=1}^{n_1}[x_{ik}(1)-\bar{x}_i(1)][x_{jk}(1)-\bar{x}_j(1)]\\&+\sum_{k=1}^{n_2}[x_{ik}(2)-\bar{x}_i(2)][x_{jk}(2)-\bar{x}_j(2)]\end{aligned}$$

由(2-4-4)式有

$$\begin{aligned}\frac{\partial U}{\partial c_i}&=2(c_1d_1+c_2d_2+\cdots+c_md_m)d_i\\&=2\Big(\sum_{j=1}^{m}c_jd_j\Big)d_i\end{aligned}$$

由（2-4-5）式有

$$\frac{\partial V}{\partial c_i}=2\sum_{j=1}^{m}c_j s_{ij}$$

前已述及

$$\frac{1}{E}=\frac{\partial U}{\partial c_i}\bigg/\frac{\partial V}{\partial c_i}$$

那么

$$\frac{2}{E}\left(\sum_{j=1}^{m}c_j d_j\right)d_i=2\left(\sum_{j=1}^{m}c_j s_{ij}\right)$$

如果令

$$p=\frac{\sum_{j=1}^{m}c_j d_j}{E}$$

则有

$$\sum_{j=1}^{m}c_j s_{ij}=pd_i \qquad (i=1,2,\cdots,m) \tag{2-4-6}$$

（2-4-6）式是一个由m个方程组成的线性方程组，其中的p是个常数，对线性方程组只起扩大或缩小p倍的作用，而并不影响待定系数c_j之间的相对比例关系，对判别函数来讲并没有什么影响，因而可令$p=1$，所以有

$$\sum_{j=1}^{m}c_j s_{ij}=d_i \qquad (i=1,\ 2,\ \cdots,\ m) \tag{2-4-7}$$

将（2-4-7）式展开，则为如下线性方程组

$$\begin{cases} s_{11}c_1+s_{12}c_2+\cdots+s_{1m}c_m=d_1 \\ s_{21}c_1+s_{22}c_2+\cdots+s_{2m}c_m=d_2 \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots \\ s_{m1}c_1+s_{m2}c_2+\cdots+s_{mm}c_m=d_m \end{cases} \tag{2-4-8}$$

（2-4-7）式也可写成矩阵形式

$$\begin{pmatrix} s_{11}s_{12} & \cdots & s_{1m} \\ s_{21}s_{22} & \cdots & s_{2m} \\ \cdots\ \cdots & \cdots & \cdots \\ s_{m1}s_{m2} & \cdots & s_{mm} \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ \cdots \\ c_m \end{pmatrix}=\begin{pmatrix} d_1 \\ d_2 \\ \cdots \\ d_m \end{pmatrix} \tag{2-4-9}$$

因而，可求解出待定系数

$$C=S^{-1}D=\begin{vmatrix} \sum_{i=1}^{m}s_{i1}^{-1}d_i \\ \sum_{i=1}^{m}s_{i2}^{-1}d_i \\ \cdots\quad\cdots \\ \sum_{i=1}^{m}s_{im}^{-1}d_i \end{vmatrix} \tag{2-4-10}$$

(2-4-10)式中的s_{ij}^{-1}为S的逆矩阵S^{-1}中的元素，解得的待定系数c_1，c_2，…，c_m回代到(2-4-2)式则可得到线性判别方程式

$$y=c_1x_1+c_2x_2+\cdots+c_mx_m$$

二、判别指标及其显著性检验

由上面求得的线性判别方程可以计算

$$\bar{y}(1)=\sum_{i=1}^{m}c_i\bar{x}_i(1)$$

$$\bar{y}(2)=\sum_{i=1}^{m}c_i\bar{x}_i(2)$$

由于$\bar{y}(1)$可由n_1组（每组有m个数据）观测值求得；同样$\bar{y}(2)$可由n_2组观测值求得，因而可取$\bar{y}(1)$与$\bar{y}(2)$的加权平均值$\bar{y}_c$作为判别指标

$$y_c=\frac{n_1\bar{y}(1)+n_2\bar{y}(2)}{n_1+n_2} \tag{2-4-11}$$

进而可用判别指标y_c作为样品类型归属的判别值。如果$y>y_c$时则判定样品属于第1类；而$y<y_c$时，则可判定样品属于第2类。

判别分析是假设样品取自不同的母体，对于两组判别分析则假定第1类及第2类是分别取自两个母体。如果对两个不同的母体，所用的m个地质变量在统计上的差异不显著时，判别显然是无意义的。这就需要对两个母体的差异进行显著性检验。

为检验显著性可用马哈拉诺比斯（Mahalanobis）距离D^2为基础构成的如下统计量F。

$$F=\left[\frac{n_1n_2}{(n_1+n_2)(n_1+n_2-2)}\right]\left[\frac{n_1+n_2-m-1}{m}\right]D^2 \tag{2-4-12}$$

式中的D^2为马哈拉诺比斯距离

$$D^2=(n_1+n_2-2)\sum_{i=1}^{m}\sum_{j=1}^{m}d_id_js_{ij}^{-1} \tag{2-4-13}$$

(2-4-13)式中的s_{ij}^{-1}为逆矩阵S^{-1}中的元素。

F服从自由度为m与(n_1+n_2-m-1)的F分布，即$F_a(m, n_1+n_2-m-1)$。m为判别使用的地质变量个数，n_1、n_2分别为母体1（第1类）、母体2（第2类）的样品数。最后查F分布表确定判别函数的显著性。如果$F>F_a$，则认为判别方程有意义，即认为判别方程能将第1类与第2类样品分开；$F\leqslant F_a$时，则认为判别方程无意义。

三、算　例

云南省南部的某个地质凹陷中，经过勘探证实，A区是一个钾盐矿区，B区是一个钠盐矿区。此外，还有许多盐泉有待作出评价，以判别每个盐泉是属于钾盐泉还是钠盐泉。A区与B区盐泉的化验分析数据见表2-4-1。

按两组判别分析方法计算，得到如下判别函数

$$y=0.6136x_1+0.5452x_2-1.1063x_3+0.0931x_4$$

进一步可以计算判别指标

表2-4-1　A区及B区的盐泉化验分析数据表

地　　区	盐泉编号	地质变量			
		$\frac{K}{Cl}10^3(x_1)$	$\frac{Br}{Cl}10^3(x_2)$	$\frac{K}{\Sigma盐}10^3(x_3)$	$\frac{K}{Br}(x_4)$
A区（钾矿泉）	1	13.85	4.79	7.8	49.6
	2	22.31	4.67	12.31	47.8
	3	28.82	4.63	16.18	62.15
	4	15.29	3.54	7.58	43.2
	5	28.29	4.90	16.12	58.7
B区（钠矿泉）	8	2.18	1.06	1.22	20.6
	10	3.85	0.8	4.06	47.1
	11	11.4	0	3.50	0
	29	3.66	2.42	2.14	15.1
	30	12.10	0	5.68	0

$$\bar{y}(A)=c_1\bar{x}_1(A)+c_2\bar{x}_2(A)+c_3\bar{x}_3(A)+c_4\bar{x}_4(A)$$
$$=0.6136\times21.712+0.5452\times4.106-1.1063\times11.998$$
$$+0.0931\times52.290=7.1566$$
$$\bar{y}(B)=c_1\bar{x}_1(B)+c_2\bar{x}_2(B)+c_3\bar{x}_3(B)+c_4\bar{x}_4(B)$$
$$=0.6136\times6.638+0.5452\times0.856-1.1063\times3.32$$
$$+0.0931\times16.56=2.4088$$

由此得到

$$y_c=\frac{n_1\bar{y}(A)+n_2\bar{y}(B)}{n_1+n_2}$$
$$=\frac{5\times7.1566+5\times2.4088}{5+5}=4.7827$$

由于$\bar{y}(A)=7.1566>y_c=4.7827$，因而$y=\sum_{i=1}^{4}c_ix_i>y_c$的$\{x_1, x_2, x_3, x_4, x_5\}$应属于钾盐泉；而$y=\sum_{i=1}^{4}c_ix_i\leqslant y_c$的$\{x_1, x_2, x_3, x_4, x_5\}$应属于钠盐泉。

首先，对已知的A、B两区的盐泉作验证，计算结果见表2-4-2。

表2-4-2　A区及B区的盐泉判别验证

地　区	盐泉编号	判别值y	判别分类	原有分类	判　别　正　误
A区（钾盐区）	1	6.008	A	A	正确
	2	7.061	A	A	正确
	3	8.094	A	A	正确
	4	6.945	A	A	正确
	5	7.968	A	A	正确

续表

地　区	盐泉编号	判别值y	判别分类	原有分类	判　别　正　误
B区 （钠盐区）	8	2.484	B	B	正确
	10	2.692	B	B	正确
	11	3.125	B	B	正确
	29	2.603	B	B	正确
	30	1.141	B	B	正确

由表2-4-2可以看出，A区的5个盐泉其判别值$y(A)$均大于判别指标y_c(4.7827)；B区的5个盐泉，其判别值$y(B)$均小于判别指标y_c。可见，由判别函数对两类盐泉的判别分类与原有分类完全一致，即对10个盐泉的判别完全正确。

现有8个未知属性的盐泉，其化验分析数据以及用判别函数计算的判别值见表2-4-3。

表2-4-3　未知盐泉的化验分析数据及判别结果

盐泉编号	地　质　变　量				判　别　结　果	
	$\frac{K}{Cl}10^3(x_1)$	$\frac{Br}{Cl}10^2(x_2)$	$\frac{K}{\Sigma盐}(x_3)$	$\frac{K}{Br}(x_4)$	判别值y	判别分类
18	8.85	3.58	5.17	26.1	3.984	B
22	28.6	2.40	1.20	127.0	29.355	A
24	20.7	6.70	7.60	30.8	10.814	A
27	7.90	2.40	4.30	33.2	4.490	B
51	3.19	3.20	1.43	9.9	3.042	B
53	12.40	5.10	4.43	24.6	7.724	A
57	16.80	5.40	2.31	31.3	13.611	A
69	15.0	2.70	5.02	64	11.082	A

下面对判别方程进行显著性检验，马哈拉诺比斯距离为

$$D^2=(n_1+n_2-2)(c_1d_1+c_2d_2+c_3d_3+c_4d_4)$$
$$=37.9744$$

由D^2可以得到检验统计量

$$F=\left[\frac{n_1n_2}{(n_1+n_2)(n_1+n_2-2)}\right]\left[\frac{n_1+n_2-m-1}{m}\right]D^2$$

$$=\left[\frac{5\times5}{(5+5)(5+5-2)}\right]\left[\frac{5+5-4-1}{4}\right]37.9744$$

$$=14.822$$

查F分布临界值表得到

$$F_\alpha(m,n_1+n_2-m-1)=F_\alpha(4,5)=\begin{cases}F_{0.01}(4,5)=11.4\\F_{0.05}(4,5)=5.19\end{cases}$$

由于$F=14.822>F_{0.01}(4,5)=11.4$，因此，可以认为$A$、$B$两组变量的平均值高度显著，说明上述判别结果是有效的。

第二节　多组判别分析

在地质研究工作中，经常会碰到样品的类型归属不仅仅限于两种类型的问题。例如，一口探井经过多种测井方法测量后，得到多种测井曲线，每种测井曲线可以看作是一个地质变量，每个层段可看作是一个样品。剖面中的某个层段可能是油层、气层、水层中的一种，也可能是干层。这种情况下，样品的归属类型为4种。又如，在石油资源评价中的含油气有利地带的概念也属于多种类型归属问题，一般可分为最有利勘探地带、有利勘探地带、中等有利勘探地带、不利勘探地带以及最不利勘探地带。此时，如果以勘探目标为样品，则样品的归属类型为5种。可见，多组判别分析在地质研究工作中更有实用价值，前面讲的两组判别分析可以看作是多组判别分析的特例。

一、多组判别方程

如果有n个样品，每个样品有m个地质变量，n个样品分别归属于h（$h>2$）种类型，每个类中的样品数不一定相等，假设其中第i类的样品数为$n_i(i=1,2,\cdots,h)$个。此时，每个观测数据应有3个下标，因而，应写成如下形式

$$x_{jki}\quad (j=1,2,\cdots,m;\ k=1,2,\cdots,n_i;\ i=1,2\cdots,h)$$

其中下标j表示第j个地质变量，下标k表示第k个样品；下标i表示第i种类型。因此，第1类观测数据应表示为

$$\begin{matrix} x_{111}, & x_{211}, & \cdots, & x_{m11} \\ x_{121}, & x_{221}, & \cdots, & x_{m21} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n_11}, & x_{2n_11}, & \cdots, & x_{mn_11} \end{matrix}$$

第2类观测数据应表示为

$$\begin{matrix} x_{112}, & x_{212}, & \cdots, & x_{m12} \\ x_{122}, & x_{222}, & \cdots, & x_{m22} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n_22}, & x_{2n_22}, & \cdots, & x_{mn_22} \\ \cdots & \cdots & \cdots & \cdots \end{matrix}$$

第h类观测数据应表示为

$$\begin{matrix} x_{11h}, & x_{21h}, & \cdots, & x_{m1h} \\ x_{12h}, & x_{22h}, & \cdots, & x_{m2h} \\ \cdots & \cdots & \cdots & \cdots \\ x_{1n_hh}, & x_{2n_hh}, & \cdots, & x_{mn_hh} \end{matrix}$$

两组判别分析，是用一个判别函数把空间分成两个域；三组判别分析需要三个判别函数把空间划分为三个域；而四组判别分析，就需要六个判别函数，如此类推。随着判别类型的增加，判别函数的个数将迅速增加。可见，对于多组判别分析，计算组间的判别函数是很不方便的。因此，多组判别分析一般采用的是计算每个样品属于各组的概率，也就是说，对于

一个归属类型尚未确知新样品，在判别它属于h个已知类型的哪一个类型时，是要计算它属于各种类型的概率p_i（$i=1,2,\cdots,h$），然后比较$p_1,p_2,\cdots,p_h$的大小，并将这个样品归入概率最大的那个类型中。所以，对于多组判别分析来说，关键是要给出一个计算p_i的算法公式。

前已述及，x_{iki}表示第i种类型，第k个样品的第j个变量。假设各类的样品都是来自不同母体的互相独立的正态分布的随机变量，即，（x_{1ki}，x_{2ki}，…，x_{mki}）$\sim N(a_i,\sigma_i)$，其中的a_i为第i类m个变量的数学期望，σ_i为第i类m个变量的协方差矩阵。为了计算上的方便，这里假定各类的协方差矩阵是相等的，即有$\sigma_1=\sigma_2=\cdots=\sigma_h$，而各类的不同之处仅仅是$a_i$（$i=1,2,\cdots,h$）。

基于上述假设，可以根据各已知归属类的样品数$n_1,n_2,\cdots,n_h$来估计各类的a及σ。

$$\hat{a}_i=\bar{x}_i=\begin{bmatrix} x_{1\cdot i} \\ x_{2\cdot i} \\ \cdots \\ x_{m\cdot i} \end{bmatrix} \quad (i=1,2,\cdots,h)$$

$$\hat{\sigma}=\frac{1}{n-m}\sum_{i=1}^{h}S_i=D$$

其中

$$S_i=[s_{ab}^{(i)}]_{m\times m}$$

$$s_{ab}^{(i)}=\sum_{k=1}^{n_i}(x_{aki}-x_{a\cdot i})(x_{bki}-x_{b\cdot i})$$

把这些估计值代入各类的密度分布表达式中，就得到各类的m个变量的联合密度分布

$$p_i(x_1,x_2,\cdots,x_m)=\frac{|D|^{-\frac{1}{2}}}{\sqrt{(2\pi)^m}}\exp\left[-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(x_a^{(i)}-x_{a\cdot i})(x_b^{(i)}-x_{b\cdot i})\right]$$

式中的D为总协方差矩阵，d_{ab}^{-1}为D的逆矩阵D^{-1}中的第a行第b列上的元素。

如果任意给定一个新样品，其m个变量的取值为$y_1,y_2,\cdots,y_m$。假定这一样品来自各类母体的可能性是相等的，则可由贝叶斯（Bayes）公式来计算这一样品来自第i类的后验概率p_i

$$p_i=\frac{q_i p_i(y_1,y_2,\cdots,y_m)}{\sum_{i=1}^{h}q_i p_i(y_1,y_2,\cdots,y_m)} \qquad (2\text{-}4\text{-}14)$$

式中，$p_i(y_1,y_2,\cdots,y_m)$为样品$y(y_1,y_2,\cdots,y_m)$属于第i组的概率密度。q_i为第i组的先验概率，可用样品的频率作为先验概率的估计值，即 $q_i=\frac{n_i}{n}$。

前面已经约定，判别样品归属类型的准则是$p_i(y_1,y_2,\cdots,y_m)$为最大，所以当判别函数

$$q_s p_s(y_1,y_2,\cdots,y_m)=\max_{1\leq i\leq h}\{q_i p_i(y_1,y_2,\cdots,y_m)\}$$

时，则把新样品$y(y_1,y_2,\cdots,y_m)$划入第s类，可见，在多组判别时，是要找出判别函数$q_i p_i$为最大的s，s即为样品归属类型的标号。

按（2-4-14）式，计算p_i的表达式应为

$$p_i=\frac{q_i\dfrac{|D|^{-\frac{1}{2}}}{(\sqrt{2\pi})^m}\exp\left[-\dfrac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(y_a-x_{a\cdot i})(y_b-x_{b\cdot i})\right]}{\sum_{i=1}^{k}q_i\dfrac{|D|^{-\frac{1}{2}}}{(\sqrt{2\pi})^m}\exp\left[-\dfrac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(y_a-x_{a\cdot i})(y_b-x_{b\cdot i})\right]}$$

$$=\frac{q_i\exp\left[-\dfrac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(y_a-x_{a\cdot i})(y_b-x_{b\cdot i})\right]}{\sum_{i=1}^{k}q_i\exp\left[-\dfrac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(y_a-x_{a\cdot i})(y_b-x_{b\cdot i})\right]}$$

对上式的分子取对数，则有

$$\ln(q_ip_i)=\ln q_i-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(y_a-x_{a\cdot i})(y_b-x_{b\cdot i})$$

$$=\ln q_i-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}(y_ay_b-x_{a\cdot i}y_b-x_{b\cdot i}y_a+x_{a\cdot i}x_{b\cdot i})$$

$$=\ln q_i-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}x_{a\cdot i}x_{b\cdot i}+\sum_{a=1}^{m}y_a\left(\sum_{b=1}^{m}d_{ab}^{-1}x_{b\cdot i}\right)$$

$$-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}y_ay_b$$

令

$$c_{0i}=-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}x_{a\cdot i}x_{b\cdot i}$$

$$c_{ai}=\sum_{b=1}^{m}d_{ab}^{-1}x_{b\cdot i}$$

那么 $$\ln(p_iq_i)=\ln q_i+c_{0i}+\sum_{a=1}^{m}y_ac_{ai}-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}y_ay_b$$

再令 $$f_i(y_1,y_2,\cdots,y_m)=\ln q_i+c_{0i}+\sum_{a=1}^{m}y_ac_{ai}\qquad(2\text{-}4\text{-}15)$$

则有 $$q_ip_i=\exp\left(f_i-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}y_ay_b\right)$$

因而

$$p_i=\frac{\exp\left(f_i-\dfrac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}y_ay_b\right)}{\sum_{i=1}^{k}\exp\left(f_i-\dfrac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}y_ay_b\right)}$$

$$=\frac{\exp(f_i)}{\sum_{i=1}^{k}\exp(f_i)}\qquad(2\text{-}4\text{-}16)$$

由（2-4-16）式可知，使f为最大的s共p_i也必然最大。所以，只要把样品$y(y_1,y_2,\cdots,y_m)$

代入到（2-4-15）式中，分别计算出$f_1(y_1,y_2,\cdots,y_m)$、$f_2(y_1,y_2,\cdots,y_m)$、…、$f_h(y_1,y_2,\cdots,y_m)$，当

$$f_s(y_1,y_2,\cdots,y_m)=\max_{1\le i\le h}\{f_i(y_1,y_2,\cdots,y_m)\}$$

时，则把样品y划归入第s类。这里称f_i为多组判别函数。

二、多组判别分析的计算步骤

多组判别分析的具体计算过程可以归结为如下步骤。

1. 计算各类的变量平均值

$$x_{j,i}=\frac{1}{n_i}\sum_{k=1}^{n_i}x_{jki} \qquad (j=1,2,\cdots,m;\ i=1,2,\cdots,h) \tag{2-4-17}$$

2. 计算各类的离差矩阵

矩阵S_i中的a行b列元素$S_{ab}^{(i)}$为

$$S_{ab}^{(i)}=\sum_{k=1}^{n_i}(x_{aki}-x_{a,i})(x_{bki}-x_{b,i}) \tag{2-4-18}$$

$$(a,b=1,2,\cdots,m)$$

那么

$$S_i=[S_{ab}^{(i)}]_{m\times m} \qquad (i=1,2,\cdots,h)$$

3. 计算总协方差矩阵

$$D=\frac{\sum_{i=1}^{h}S_i}{\sum_{i=1}^{h}n_i-h}=\frac{\sum_{i=1}^{h}S_i}{n-h} \tag{2-4-19}$$

4. 求D的逆矩阵D^{-1}

$$D^{-1}=[d_{ab}^{-1}]$$

5. 计算各组的判别函数

如果（$d_{a1}^{-1},d_{a2}^{-1},\cdots,d_{am}^{-1}$）是$D^{-1}$的第$a$行，需要计算

$$c_{ai}=\sum_{b=1}^{m}d_{ab}^{-1}x_{b,i} \qquad (i=1,2,\cdots,h)$$

$$c_{0i}=-\frac{1}{2}\sum_{a=1}^{m}\sum_{b=1}^{m}d_{ab}^{-1}x_{a,i}x_{b,i} \qquad (i=1,2,\cdots,h)$$

再由c_{ai}、c_{0i}确定第i组的判别函数f_i

$$f_i=\ln q_i+c_{0i}+\sum_{a=1}^{m}y_ac_{ai} \qquad (i=1,2,\cdots,h)$$

6. 计算新样品属于第i组的概率p_i

$$p_i=\frac{\exp(f_i)}{\sum_{i=1}^{h}\exp(f_i)}$$

7. 计算已知分类样品的分类矩阵B_0

$$B_0=[b_{ij}]_{h\times h}$$

矩阵$\hat{B}$。中的元素b_{ij}表示已知属于第i组的n_i个样品，经过计算p_i后重新分组时，属于第j组的样品个数。

如果计算结果$b_{ii}=n_i(i=1,2,\cdots,h)$，那就表示计算后的分类与计算前的分类是完全一致的。

三、判别函数的检验

在实际地质研究工作中，如果有n个样品已知分别属于h个类型，通常在地质概念上类与类之间的差别是明确的。但是，经过多组判别分析的分类可能与原来的分类有出入，甚至会有较大的出入。出现这种情况的原因可能是所选取的地质变量不能充分表现各类之间的差别。因此，需要检验所选取的m个地质变量是否有能力区分开这h个类。在此可用马哈拉诺比斯距离D^2作为统计量进行检验。

为求得D^2，先要计算各个地质变量的总平均值$x_{j..}$。

$$x_{j..}=\frac{1}{n}\sum_{i=1}^{h}\sum_{k=1}^{n_i}x_{ikj}=\frac{\sum_{i=1}^{h}n_i x_{j.i}}{\sum_{i=1}^{h}n_i} \qquad (2\text{-}4\text{-}20)$$

$$(j=1,2,\cdots,m)$$

再根据已求得的逆矩阵D^{-1}计算D^2。

$$D^2=\sum_{a=1}^{m}\sum_{b=1}^{m}\sum_{i=1}^{h}n_i d_{ab}^{-1}(x_{a.i}-x_{a..})(x_{b.i}-x_{b..}) \qquad (2\text{-}4\text{-}21)$$

统计量D^2服从自由度为$p(h-1)$的x^2分布，查表可定出D^2的临界值。当算得的D^2大于临界值时，可以断定这m个地质变量可以判别这h个类型的归属问题；否则，就表明这m个地质变量还不足以鉴别这h个类型的划分问题。在这种情况下，应剔除一些不重要的地质变量或者引进一些新的有效的地质变量。

四、算　　例

某含盐矿区有钾盐泉（A类）、盐岩盐泉（B类）、和钾矿化盐泉（C类）三类矿泉。此外，还有一批未知分类的盐泉。现应用三组判别分析对未知盐泉进行分类，以指导钾盐的找矿工作。已知分类的A、B、C矿区盐泉的水化学分析资料见表2-4-4。未知分类的待判盐泉的水化学分析资料见表2-4-5。

表2-4-4　已知分类盐泉的水化学分析资料

盐泉分类	序号	地质变量						
		矿化度(g/L)	$\frac{Br}{Cl}10^3$	$\frac{K}{\Sigma 盐}10^3$	$\frac{K}{Cl}10^3$	$\frac{Na}{K}$	$\frac{Mg}{Cl}10^2$	$\frac{eNa}{eCl}$
钾	1	11.853	0.48	14.36	25.21	25.21	0.81	0.98
	2	45.596	0.526	13.85	24.04	26.01	0.91	0.96
	3	3.525	0.086	24.4	49.3	11.30	6.82	0.85
盐	4	3.681	0.327	13.57	25.12	26.00	0.82	1.01
	5	48.287	0.386	14.5	25.9	23.32	2.18	0.93
	6	254.643	0.43	14.5	24.7	25.41	0.41	0.96
泉	7	17.956	0.28	9.75	17.05	37.20	0.464	0.93
	8	7.37	0.506	13.6	34.21	10.69	8.8	0.56
	9	310.748	0.493	10.94	18.31	33.68	0.667	0.94

盐泉分类	序号	地质变量						
		矿化度(g/L)	$\frac{Br}{Cl}10^3$	$\frac{K}{\Sigma盐}10^3$	$\frac{K}{Cl}10^3$	$\frac{Na}{K}$	$\frac{Mg}{Cl}10^2$	$\frac{\varepsilon Na}{\varepsilon Cl}$
钾盐泉	10	314.152	0.191	14.65	24.5	24.72	0.672	0.93
	11	183.01	0.296	8.75	14.92	42.25	0.274	0.97
	12	6.742	0.19	5.93	12.01	57.95	0.99	1.07
钾矿化盐泉	13	4.741	0.14	6.9	15.7	39.2	3.24	0.95
	14	4.223	0.34	3.8	7.1	88.2	1.11	0.97
	15	6.442	0.19	4.7	9.1	23.2	0.74	1.08
	16	16.234	0.39	3.4	5.4	121.5	0.42	1.00
	17	10.585	0.42	2.4	4.7	135.6	0.87	0.98
	18	39.416	0.32	2.8	5.1	129.3	0.45	1.01
	19	37.228	0.26	3.0	5.6	115.6	0.90	1.00
	20	23.535	0.23	2.6	4.6	141.2	0.31	1.02
	21	5.398	-0.12	2.8	6.2	111.1	1.14	1.07
	22	92.589	0.26	2.7	4.8	135.6	0.26	1.00
	23	145.228	0.30	2.7	4.7	135.4	0.24	0.99
	24	43.865	0.20	2.3	4.0	161.6	0.27	1.01
盐岩盐泉	25	48.621	0.082	2.057	3.847	170.15	0.94	1.00
	26	288.149	0.148	1.763	2.968	215.83	0.14	0.98
	27	316.604	0.317	1.453	2.432	263.41	0.249	0.98
	28	307.31	0.173	1.627	2.729	235.70	0.214	0.99
	29	82.17	0.105	1.217	2.188	297.79	0.33	1.00
	30	322.515	0.312	1.382	2.320	282.21	0.024	1.00
	31	31.409	0.145	0.859	1.567	407.34	0.726	0.98
	32	78.938	0.033	0.97	1.687	282.50	0.244	0.99
	33	106.281	0.053	0.941	1.658	391.05	0.270	1.00
	34	256.58	0.297	0.899	1.476	410.30	0.239	0.93
	35	304.092	0.283	0.789	1.357	483.36	0.193	1.01
	36	240.446	0.042	0.741	1.266	500.77	0.29	0.99

表2-4-5 待判盐泉的水化分析资料

序号	地质变量						
	矿化度(g/L)	$\frac{Br}{Cl}10^3$	$\frac{K}{\Sigma盐}10^3$	$\frac{K}{Cl}10^3$	$\frac{Na}{K}$	$\frac{Mg}{Cl}10^2$	$\frac{\varepsilon Na}{\varepsilon Cl}$
1	3.777	0.87	15.4	28.2	7.6	0.40	0.77
2	14.829	0.55	19.3	33.7	18.0	0.93	0.96
3	9.532	0.45	17.0	28.2	13.7	0.04	0.60
4	1.960	0	25.5	66.3	10.6	0.87	1.08
5	2.254	0.35	6.7	15.1	42.9	2.08	1.00
6	11.218	0.28	6.2	12.3	51.6	1.23	0.99
7	7.437	0.57	3.4	7.8	91.3	1.27	1.10

续表

序号	地质变量						
	矿化度 (g/L)	$\frac{Br}{Cl}10^3$	$\frac{K}{\Sigma盐}10^3$	$\frac{K}{Cl}10^3$	$\frac{Na}{K}$	$\frac{Mg}{Cl}10^2$	$\frac{\varepsilon Na}{\varepsilon Cl}$
8	5.825	0.48	3.8	7.5	87.9	1.02	1.01
9	12.430	0.23	5.6	10.6	61.4	0.75	1.00
10	62.856	0.34	5.2	9.0	7[illegible].2	0.50	0.99
11	1.128	0.29	7.5	17.7	34.0	5.12	0.93
12	39.160	0.26	5.4	9.6	65.4	0.46	1.60
13	6.014	0.11	5.7	12.7	46.4	0.24	0.91
14	19.388	0.07	6.7	13.3	42.8	0.86	0.99
15	8.886	0.40	3.9	8.3	86.1	1.10	1.10
16	2.391	0.39	4.2	10.4	70.0	2.76	1.12
17	18.322	0.21	4.4	8.8	74.1	1.22	1.00
18	5.452	0.42	2.3	4.3	15.62	0.73	1.03
19	1.133	0	9.7	21.2	24.3	5.64	0.80
20	3.299	0.18	3.0	5.2	103.5	2.67	0.82
21	25.431	0.37	2.0	3.8	173.2	0.51	1.01
22	1.366	0	3.3	8.3	75.8	2.23	0.99
23	301.351	0.41	3.3	5.8	114.4	0.29	0.98
24	11.948	0.38	1.8	3.4	199.2	0.31	1.05
25	68.224	0.24	2.4	4.3	145.8	0.83	0.97
26	37.864	0.30	1.9	3.8	169.9	0.67	0.99
27	19.283	0.12	2.3	4.3	146.2	1.27	0.98
28	17.818	0.38	1.7	3.1	219.7	0.33	1.04
29	5.895	0.25	1.7	3.7	182.0	1.16	1.05
30	21.770	0.25	1.7	3.3	196.7	0.41	1.01
31	3.434	0.10	2.0	4.7	165.7	0.75	1.20
32	10.038	0.15	1.9	3.7	179.6	0.74	1.01
33	87.393	0.32	1.8	3.2	197.4	0.55	0.97
34	2.771	0	2.0	4.0	158.6	1.33	0.98
35	6.033	0.26	1.4	2.6	239.4	1.08	0.94
36	9.386	0.51	3.5	7.1	88.9	1.30	0.97

经过三组判别分析计算，已知分类盐泉的判别检验结果见表2-4-6，对未知分类盐泉的判别结果见表2-4-7。在上述属于A类及C类的盐泉所在地区均有可能找到钾盐矿。

表2-4-6 已知分类盐泉的判别检验

盐泉分类	序号	各类判别函数值			原分类	判别分类	判别正误
		f_1	f_2	f_3			
钾盐泉	1	648.022	629.975	615.291	A	A	正确
	2	660.225	644.850	630.492	A	A	正确
	3	698.348	676.643	665.759	A	A	正确
	4	625.036	604.776	591.659	A	A	正确

续表

盐泉分类	序号	各类判别函数值			原分类	判别分类	判别正误
		f_1	f_2	f_3			
钾盐泉	5	681.548	672.737	660.021	A	A	正确
	6	633.192	614.307	605.918	A	A	正确
	7	598.091	589.693	579.596	A	A	正确
	8	586.212	575.790	567.314	A	A	正确
	9	681.710	675.073	667.722	A	A	正确
	10	583.764	568.033	564.983	A	A	正确
	11	614.722	609.144	602.746	A	A	正确
	12	707.026	712.616	702.770	A	C	错误
钾矿化盐泉	13	691.891	701.049	692.638	C	C	正确
	14	680.468	692.250	683.317	C	C	正确
	15	665.065	672.462	664.292	C	C	正确
	16	686.761	697.416	689.496	C	C	正确
	17	699.820	713.185	705.181	C	C	正确
	18	683.701	695.264	688.641	C	C	正确
	19	691.620	706.171	699.668	C	C	正确
	20	664.480	675.562	671.307	C	C	正确
	21	732.885	750.527	743.740	C	C	正确
	22	652.246	662.346	658.408	C	C	正确
	23	659.061	669.006	665.753	C	C	正确
	24	649.256	661.590	658.278	C	C	正确
盐岩盐泉	25	664.102	682.550	681.076	B	C	错误
	26	663.341	645.570	650.949	B	B	正确
	27	693.304	707.511	712.684	B	B	正确
	28	663.602	677.809	683.701	B	B	正确
	29	634.920	650.190	654.640	B	B	正确
	30	697.494	710.786	716.629	B	B	正确
	31	992.958	1060.271	1058.629	B	B	错误
	32	603.870	618.679	627.742	B	C	正确
	33	627.302	643.172	652.522	B	B	正确
	34	626.651	638.662	649.767	B	B	正确
	35	724.765	741.881	755.045	B	B	正确
	36	628.078	644.747	653.384	B	B	正确

表2-4-7　对未知盐泉的判别

序　　号	各类判别函数值			max(f)	判别分类
	f_1	f_2	f_3		
1	412.295	365.947	351.544	f_1	A
2	628.211	593.282	581.017	f_1	A
3	142.516	91.267	86.496	f_1	A
4	−18.626	−170.165	−169.977	f_1	A
5	705.960	707.361	695.818	f_2	C
6	658.418	661.262	651.599	f_2	C
7	871.386	887.382	872.260	f_2	C
8	743.590	754.867	743.066	f_2	C
9	636.971	640.317	631.851	f_2	C
10	664.964	670.742	662.141	f_2	C
11	844.425	866.020	853.683	f_2	C
12	641.697	645.868	637.653	f_2	C
13	396.204	377.777	374.605	f_1	A
14	568.189	569.315	559.407	f_1	A
15	822.102	837.045	824.300	f_2	C
16	946.923	972.441	957.349	f_2	C
17	674.672	684.975	667.141	f_2	C
18	748.448	765.033	751.095	f_2	C
19	660.572	675.572	669.638	f_2	C
20	647.266	672.266	667.890	f_2	C
21	702.830	716.604	710.671	f_2	C
22	2492.316	2723.480	2698.264	f_2	C
23	702.383	712.008	709.492	f_2	C
24	736.890	751.938	745.734	f_2	C
25	660.723	675.605	672.005	f_2	C
26	673.567	687.266	682.982	f_2	C
27	671.592	691.248	687.798	f_2	C
28	732.519	748.064	742.940	f_2	C
29	767.943	790.266	784.495	f_2	C
30	665.843	679.380	676.279	f_2	C
31	858.669	883.675	876.253	f_2	C
32	666.338	682.750	679.884	f_2	C
33	664.774	678.513	676.427	f_2	C
34	641.059	661.123	650.735	f_2	C
35	650.706	668.434	667.43	f_2	C
36	727.021	738.789	727.480	f_2	C

第三节　逐步判别分析

在实际研究工作中，研究人员总希望能用最少的变量就能解决需要判别的问题，也就是说，应当选择少数的有效变量进行判别分析；多余的变量参加判别时，不仅会使计算工作量

加大，还有可能由于增加变量而产生变量间的相关性，造成计算上的困难。因而，自然会产生类似逐步回归的想法，即对变量按其对判别分类的重要性，在计算过程中有进有出，保留那些对判别类型起主要作用的变量，剔除那些对判别类型不起作用或者起作用不大的变量。

一、逐步判别方法原理

如果有n个样品，每个样品分析了m个变量，n个样品分别归属于h个类，其中第i类的样品数为$n_i(i=1,2,\cdots,h)$。这里假定各类的样品都是来自不同母体、相互独立的、正态分布的随机向量，即

$$X_{k1}(k=1,2,\cdots,n_1)\sim N(a_1,\sigma)$$
$$X_{k2}(k=1,2,\cdots,n_2)\sim N(a_2,\sigma)$$
$$\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots$$
$$X_{kh}(k=1,2,\cdots,n_h)\sim N(a_h,\sigma)$$

而其中的

$$X_{ki}=[x_{1ki},\ x_{2ki},\cdots,x_{mki}]'$$

令

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{h}\sum_{k=1}^{n_i}X_{ki}$$

$$\overline{X}_i=\frac{1}{n_i}\sum_{i=1}^{n_i}X_{ki}$$

其中

$$n=\sum_{i=1}^{h}n_i$$

$\overline{X}$、$\overline{X}_i$分别为总平均向量和分组平均向量。若以W、V分别表示组内离差矩阵和总离差矩阵，则有

$$W=\sum_{i=1}^{h}\sum_{k=1}^{n_i}(X_{ki}-\overline{X}_i)(X_{ki}-\overline{X}_i)'$$

$$V=\sum_{i=1}^{h}\sum_{k=1}^{n_i}(X_{ki}-\overline{X})(X_{ki}-\overline{X})'$$

假如已经计算了t步，当时已选入了p个变量$x_1,x_2,\cdots,x_p$进入判别函数。再引进第$(p+1)$个变量x_r时，为考察该变量的判别效果，可用F_{1r}作为检验统计量

$$F_{1r}=\frac{1-A_r}{A_r}\cdot\frac{n-p-h}{h-1}=\frac{v_{rr}^{(t)}-w_{rr}^{(t)}}{w_{rr}^{(t)}}\cdot\frac{n-p-h}{h-1} \qquad (2\text{-}4\text{-}22)$$

F_{1r}服从$F(h-1,n-p-h)$分布，其中

$$A_r=\frac{w_{rr}^{(t)}}{v_{rr}^{(t)}}$$

(2-4-22)式中的$w_{rr}^{(t)}$、$v_{rr}^{(t)}$分别是计算到第t步时矩阵W、V中r行r列的元素。当

$$F_{1r}>F_a(h-1,n-p-h)$$

时，则认为变量x_r的判别效果显著，应把变量x_r引入到判别函数之中。

逐步判别分析的计算过程与逐步回归分析的计算过程很相似，即变量有进有出，如果计

算了t步已引进了p个变量，一般情况下引进的变量数少于或等于计算步数，即$p\leqslant t$。为了便于讨论问题，假定变量x_r是在第t步引入判别函数的，即前$t-1$步计算后已引进了（$p-1$）个变量，其中恰好不包括变量x_r。这样一来，我们要讨论的问题就归结为考察第t步引入的变量x_r的判别效果问题。此时有

$$A_r=\frac{w_{rr}^{(t-1)}}{v_{rr}^{(t-1)}}$$

逐步判别与逐步回归计算过程相似，每引入一个变量或剔除一个变量时，W矩阵和V矩阵都要进行一次变换，即

$$w_{ij}^{(t)}=\begin{cases} w_{rj}^{(t-1)}/w_{rr}^{(t-1)} & (i=r;\ j\neq r) \\ w_{ij}^{(t-1)}-w_{ir}^{(t-1)}\cdot w_{rj}^{(t-1)}/w_{rr}^{(t-1)} & (i\neq r;\ j\neq r) \\ -w_{ir}^{(t-1)}/w_{rr}^{(t-1)} & (i\neq r;\ j=r) \\ 1/w_{rr}^{(t-1)} & (i=r;\ j=r) \end{cases} \tag{2-4-23}$$

$$v_{ij}^{(t)}=\begin{cases} v_{rj}^{(t-1)}/v_{rr}^{(t-1)} & (i=r;\ j\neq r) \\ v_{ij}^{(t-1)}-v_{ir}^{(t-1)}\cdot v_{rj}^{(t-1)}/v_{rr}^{(t-1)} & (i\neq r;\ j\neq r) \\ -v_{ir}^{(t-1)}/v_{rr}^{(t-1)} & (i\neq r;\ j=r) \\ 1/v_{rr}^{(t-1)} & (i=r;\ j=r) \end{cases} \tag{2-4-24}$$

那么

$$A_r=\frac{w_{rr}^{(t-1)}}{v_{rr}^{(t-1)}}=\frac{1/w_{rr}^{(t)}}{1/v_{rr}^{(t)}}=\frac{v_{rr}^{(t)}}{w_{rr}^{(t)}}$$

由此，可以建立剔除变量x_r的检验统计量F_{2r}

$$F_{2r}=\frac{1-A_r}{A_r}\cdot\frac{n-(p-1)-h}{h-1}=\frac{w_{rr}^{(t)}-v_{rr}^{(t)}}{v_{rr}^{(t)}}\cdot\frac{n-(p-1)-h}{h-1} \tag{2-4-25}$$

F_{2r}服从$F(h-1,n-p+1-h)$分布，当$F_{2r}\leqslant F_{\alpha}(h-1,n-p+1-h)$时，则认为变量$x_r$的判别效果不显著，可以从判别函数中剔除。

二、逐步判别分析的计算步骤

逐步判别分析的具体计算过程可以归结为如下步骤。

1. 计算原始数据的分类平均值及总平均值

如果有n个样品，每个样品有m个变量，n个样品分别属于h种类型，其中第i类有n_i个样品。设原始数据为

$$x_{jki}\ (j=1,2,\cdots,m;\ k=1,2,\cdots,n_i;\ i=1,2,\cdots,h)$$

首先计算各变量的分类平均值$x_{j\cdot i}$以及总平均值$x_{j\cdot\cdot}$

$$x_{j\cdot i}=\frac{1}{n_i}\sum_{k=1}^{n_i}x_{jki}\qquad (j=1,2,\cdots,m;\ i=1,2,\cdots,h) \tag{2-4-26}$$

$$x_{j\cdot\cdot}=\frac{1}{n}\sum_{i=1}^{h}\sum_{k=1}^{n_i}x_{jki}\qquad (j=1,2,\cdots,m) \tag{2-4-27}$$

2. 计算每个变量的类内离差矩阵W及总离差矩阵T

$$W=[w_{ab}]_{m\times m}$$

其中 $$w_{ab}=\sum_{i=1}^{h}\sum_{k=1}^{n_i}(x_{aki}-x_{a\cdot i})(x_{bki}-x_{b\cdot i}) \qquad (2\text{-}4\text{-}28)$$

$$(a,b=1,2,\cdots,m)$$

$$V=[t_{ab}]_{m\times m}$$

其中 $$v_{ab}=\sum_{i=1}^{h}\sum_{k=1}^{n_i}(x_{aki}-x_{a\cdot\cdot})(x_{bki}-x_{b\cdot\cdot}) \qquad (2\text{-}4\text{-}29)$$

$$(a,b=1,2,\cdots,m)$$

3. **引进和剔除变量**

如果已经计算了t步，在判别函数中已引进了p个变量，要对已引进的变量计算

$$A_j=\frac{v_{jj}^{(t)}}{w_{jj}^{(t)}} \qquad (j=1,2,\cdots,p)$$

根据这一计算结果，可首先考虑在已选入的变量中剔除判别效果最不显著的变量。即找出A_j值为最大的变量（A_j值最大，则F_{2j}值最小）。

$$A_r=\max_{j=1,2,\cdots,p} A_j$$

作F检验

$$F_{2r}=\frac{1-A_r}{A_r}\cdot\frac{n-p+1-h}{h-1}$$

若 $F_{2r}\leqslant F_a(h-1,n-p+1-h)$时，则把变量$x_r$从判别函数中剔除，并按（2-4-23）式、（2-4-24）式对矩阵W、V进行变换。

如果$F_{2r}>F_a(h-1,n-p+1-h)$时，则不能剔除已选入的变量，此时要对尚未引进的变量计算

$$A_i=\frac{w_{ii}^{(t)}}{v_{ii}^{(t)}} \qquad (i=1,2,\cdots,m-p)$$

并且找出判别效果最显著的变量，即找出A_i值为最小的变量（A_i值最小，则F_{1i}值最大）

$$A_r=\min_{i=1,2,\cdots,p} A_i$$

作F检验

$$F_{1r}=\frac{1-A_r}{A_r}\cdot\frac{n-p-h}{h-1}$$

若 $F_{1r}>F_a(h-1,n-p-h)$，则把变量x_r引入到判别函数中，然后按（2-4-23）式、（2-4-24）式对矩阵W、V进行变换。

至此，完成了（$t+1$）步的计算。之后仿此重复进行（$t+2$）、（$t+3$）、… 步的计算，直至既不能剔除变量也不能引进变量时，逐步判别计算过程结束。

4. **判别分类**

如果逐步判别计算总共进行了（$t+1$）步，引入p个变量进入了判别函数，则对每一个样品可以计算h个判别函数f_i（$i=1,2,\cdots,h$）。对于一个给定的样品y（$y_1,y_2,\cdots,y_m$），（y可以是已知分类的n个样品中的一个，也可以是新的未知分类的待判样品），它属于第i类的判别函数f_i为

$$f_i(y)=\ln q_i+c_{0i}+\sum_{a=1}^{p} y_a c_{ai} \tag{2-4-30}$$

式中的q_i为第i组的先验概率，可用已知样品的频数作为估计值，即

$$\hat{q}_i=\frac{n_i}{n}$$

$$c_{ai}=\sum_{b=1}^{p} d_{ab}^{-1} x_{b.i}$$

$$c_{0i}=-\frac{1}{2}\sum_{a=1}^{p}\sum_{b=1}^{p} d_{ab}^{-1} x_{a.i} x_{b.i}$$

$$=-\frac{1}{2}\sum_{a=1}^{p} c_{ai} x_{a.i}$$

由总协方差矩阵D可以导出

$$D=(d_{ab})=\frac{\sum_{i=1}^{h} S_i}{n-h}$$

$$=\frac{1}{n-h}\sum_{i=1}^{h}\sum_{k=1}^{n_i}(X_{ki}-\overline{X}_i)(X_{ki}-\overline{X}_i)'$$

$$=\frac{1}{n-h}W$$

$$=\frac{1}{n-h}(w_{ab})$$

即有

$$d_{ab}=\frac{1}{n-h}w_{ab}$$

$$d_{ab}^{-1}=(n-h)w_{ab}^{-1}$$

因而

$$c_{ai}=(n-h)\sum_{b=1}^{p} w_{ab}^{-1} x_{b.i} \quad (i=1,2,\cdots,h) \tag{2-4-31}$$

$$c_{0i}=-\frac{1}{2}\sum_{a=1}^{p} c_{ai} x_{a.i} \quad (i=1,2,\cdots,h) \tag{2-4-32}$$

由$f_i(y)$可计算样品属于第i组的后验概率

$$p_i(y)=\frac{\exp[f_i(y)]}{\sum_{i=1}^{h}\exp[f_i(y)]}$$

如果

$$p_s=\max_{1\le i\le h}[p_i(y)]$$

则把样品划入第s类。

三、算　例

为了研究我国生油岩的演化阶段，表2-4-8中列出了我国有关探区的77个生油岩的地层年龄（t）、现今地温（T）及埋藏深度（H）。这些生油岩处于未成熟、成熟、高成熟、过成熟等不同的演化阶段。

表2-4-8　我国生油岩的热演化参数表

演化阶段	序号	地区	热演化参数		
			$T+273$ (°C)	t (10^6a)	H (m)
未成熟	1	松辽盆地（青$_{2+3}$）	337	125.005	1000
	2	松辽盆地（青$_1$）	328	123	1009
	3	松辽盆地（姚$_2$）	334	125	1750
	4	岐口凹陷	347	33.75	1680
	5	泌阳凹陷	342	27.998	1460
	6	辽河盆地	357	41.5	1900
	7	东台凹陷（阜宁组）	359	32	1450
	8	东台凹陷（阜宁组）	358.5	31.8	1400
	9	湖北（二叠系）	338	242	3200
	10	潜江凹陷	343	34.44	1600
	11	高邮凹陷	348	15	1980
	12	惠民凹陷	351	8.04	1850
	13	沾化凹陷	344	10.76	1600
	14	东明凹陷	352	8.5	2250
	15	松辽盆地（姚2）	338.5	127	1480
成熟	16	松辽盆地（青$_{2+3}$）	341	127	1550
	17	松辽盆地（青$_{2+3}$）	367	127	2199
	18	松辽盆地（青$_1$）	337.5	124.15	1200
	19	松辽盆地（青）	362	124.2	1799
	20	松辽盆地（姚$_2$）	364	120.4	1851
	21	松辽盆地（姚$_2$）	370	120.4	1970
	22	岐口凹陷	359	35	2001
	23	泌阳凹陷	353	31	1700
	24	泌阳凹陷	367	31	2099
	25	辽河盆地	364	48.5	2001
	26	东台凹陷（阜宁组）	364	34	2201
	27	湖北（二叠系）	384	264	4400
	28	湖北（二叠系）	348	257.6	3450
	29	江汉盆地（潜江组）	388	36.755	2800
	30	高邮凹陷（集宁组）	356	17.5	2200
	31	高邮凹陷（集宁组）	360	18	2700
	32	沾化凹陷	367	10.7	2300
	33	沾化凹陷	364	12.8	2200
	34	沾化凹陷	368	13	2500
	35	东明凹陷	362	11.6	3500

续表

演化阶段	序号	地区	热演化参数		
			$T+273$ (°C)	t (10^6a)	H (m)
高成熟	36	松辽盆地(青$_{2+3}$)	380	129	3100
	37	松辽盆地(青$_1$)	377	127.88	2350
	38	松辽盆地(青$_1$)	402	127.68	2499
	39	松辽盆地(姚$_2$)	374	124.6	2100
	40	松辽盆地(姚$_2$)	383	124.64	2299
	41	松辽盆地(姚$_2$)	377	121.8	2150
	42	松辽盆地(姚$_2$)	389	125.008	2401
	43	松辽盆地(姚$_2$)	425	121.85	3349
	44	岐口凹陷	419	36	3701
	45	湖北(二叠系)	392	271.25	4050
	46	泌阳凹陷	376	34.01	2300
	47	泌阳凹陷	392	34.05	2799
	48	东台凹陷(阜宁组)	410	64.25	3201
	49	东台凹陷(阜宁组)	398	37.005	2900
	50	东台凹陷(阜宁组)	430	64.29	3750
	51	东台凹陷(阜宁组)	403	35.5	3701
	52	湖北(二叠系)	399	271.25	4001
	53	江汉盆地(潜江组)	422	36.5	3799
	54	高邮凹陷(阜宁组)	380	23.5	3000
	55	高邮凹陷(阜宁组)	382	24	3400
	56	惠民凹陷	383	13	2800
	57	惠民凹陷	386	13.5	3100
	58	高邮凹陷(阜宁组)	399	25	4100
	59	沾化凹陷	381	16	2700
	60	沾化凹陷	385	16.5	3000
	61	沾化凹陷	387	16.5	3100
	62	松辽盆地(姚$_2$)	430	123.205	3450
	63	东明凹陷	410	15.05	3550
	64	东明凹陷	412	18	3550
	65	松辽盆地(青$_{2+3}$)	400	130	3400
过成熟	66	松辽盆地(姚$_2$)	434	123.205	3555
	67	江汉盆地(潜江组)	440	37.005	4300
	68	东台凹陷(阜宁组)	444	60.005	4100
	69	东台凹陷(阜宁组)	433	57.005	5100
	70	湖北(二叠系)	469	285	5701
	71	湖北(二叠系)	572	285.5	7199
	72	高邮凹陷,(阜宁组)	403	26	4001
	73	东明凹陷	435	20	3700
	74	东明凹陷	422	18	3800
	75	泌阳凹陷	433	36.05	4140
	76	湖北(二叠系)	454	271.29	5200
	77	湖北(二叠系)	440	271.29	4900

用逐步判别分析计算不同演化阶段的判别函数，共设计了6个变量：$T+273$、t、H、$1/H$、$\ln(T+273)$、$1/(T+273)$。变量引入判别函数式的次序及F检验的统计量见表2-4-9。

表2-4-9　变量引入次序及F检验

引入次序	变量号	变　量	F 检 验 值
1	x_6	$1/(T+273)$	99.0647
2	x_1	$T+273$	8.12767
3	x_5	$\ln(T+273)$	4.88763
4	x_4	$1/H$	2.22838
5	x_3	H	1.86462
6	x_2	t	0.35560

经过计算，各演化阶段判别函数式如下：

（1）未成熟阶段的判别函数为：

$$F_1=\ln(15/77)+25129396.419x_6-425.989x_1+245471.857x_5+2765481.140x_4-0.038551x_3+4.49185x_2-680871$$

（2）成熟阶段的判别函数为：

$$F_2=\ln(20/77)+25125826.886x_6-426.672x_1+245777.661x_5+2778122.158x_4-0.037096x_3+4.49747x_2-682422$$

（3）高成熟阶段的判别函数为：

$$F_3=\ln(30/77)+25086170.626x_6-427.208x_1+245963.611x_5+2774853.714x_4-0.037614x_3+4.50006x_2-683215$$

（4）过成熟阶段的判别函数为：

$$F_4=\ln(12/77)+25072062.781x_6-427.276x_1+246003.870x_5+2777142.253x_4-0.036213x_3+4.499517x_2-683402$$

各生油阶段的判别函数对样品的正判率都超过了80%，见表2-4-10，因此可以用这些判别函数来确定生油岩的演化阶段。

表2-4-10　各生油阶段判别函数的正判率

演化阶段	正判率（%）	演化阶段	正判率（%）
未成熟	87	高成熟	93
成　熟	90	过成熟	83

第五章 因 子 分 析

因子分析是把一些具有错综复杂关系的因子（样品或变量）归结为数量较少的几个综合因子的一种多元统计分析方法。

在某些情况下，所研究的地质问题往往需要进行多因子的综合分析。如果这些因子是相互独立的，则可以把问题化为若干个单因素，逐一进行分析。但是，在多数情况下因子之间存在相关关系，而不便于对所研究的问题进行因素分解。因子分析是采取降维的方法，设法找出少数的几个综合因子来代表原来的众多因子，而这几个综合因子既能尽量多地保留原有众多因子的信息，同时它们彼此之间又是相互独立的。如果原来众多因子之间存在着一定的相关性，那么，必然存在着起支配作用的共同因素。因子分析正是根据这一点，从原有的因子相关矩阵出发，通过研究其内部结构，找出若干个对这些因子起着支配作用的独立的综合因子，亦即用原有因子的线性组合来表达已知的观测数据。这样，既合理地解释了原有众多因子之间的相关性，又简化了原有观测系统，同时也抓住了控制原有观测数据的主要矛盾。

从地质研究的应用角度看，因子分析有以下三个方面的用途：

1. 压缩原始数据

地质人员在收集资料时，总是希望能够得到尽可能多的地质数据；而在综合研究时，又希望以尽可能少的地质数据反映自己的观点。因子分析恰好可以提供一条科学的途径，使原始数据在数量上大大精简，以利于研究人员进行综合分析。特别需要指出的是，通过因子分析不仅能使数据量大大地压缩，而且也不会影响研究结论的可靠性。因为从成因意义上讲，被因子分析压缩后的数据在质量上提高了，压缩后的数据仍然包含着原始数据中的绝大部分成因信息。

由此可见，因子分析的第一个作用就是在不损失或少损失地质成因信息的前提下，尽可能地对原始数据进行精炼，从而有利于提高综合研究工作的水平。

2. 指示地质成因的推理方向

因子分析能够把庞杂、纷乱的原始数据按地质成因上的联系进行归纳、整理、精炼以及分类，从而可以理出几条比较客观的成因线索，这就为研究人员指出了逻辑推理方向，启发人们去思考相应的地质成因结论。

3. 分解叠加的地质过程

绝大多数情况下，地质现象都是多种成因地质过程的叠加产物，既有时间上不同过程的叠加，也有空间上不同过程的叠加。这些过程相互干扰、相互掩盖，使得每个独立过程的特征往往面貌不清，因而，给研究地质成因带来了复杂性与多解性。

显然，就地质成因的研究意义而言，地质家们感兴趣的是弄清每个单一地质过程的性质和特征，了解它们在时间上和空间上的分布和发育程度，而绝不是含混的、不清的复合地质过程。而因子分析对解决这种问题恰好可以提供一些巧妙的途径。

上述三个方面是目前应用因子分析研究地质问题的主要用途。因子分析在其他方面的潜在的解决地质问题的能力，尚需继续探索与开拓。

因子分析在沉积盆地蚀源区的研究、沉积物的粒度分析、沉积相研究、地层分析、古生态与古环境的研究、岩浆岩岩石化学成分的研究、变质岩的原岩恢复、矿床成因研究、矿物的类质同象研究、地球化学研究，水化学研究等方面均已取得不同程度的进展。

因子分析与聚类分析相仿，也分为R型与Q型两种类型。R型因子分析是研究变量之间的相关关系的方法，它是通过对变量的相关系数矩阵内部结构的研究，找出控制所有变量的几个主要成分，所以，R型因子分析又称主成分分析。而Q型因子分析是研究样品之间的相关关系的方法，它是通过对样品的相似系数矩阵内部结构的研究，找出控制所有样品的几个主要因素，所以，Q型因子分析又称主因素分析。这两种因子分析的全部运算过程实际上是一样的，只不过它们的出发点不同：R型因子分析是从相关系数矩阵出发，而Q型因子分析是从相似系数矩阵出发。对于同一批观测数据，可以根据所研究地质问题的目的来决定采用哪一种因子分析方法。

第一节 因子分析的数学模型

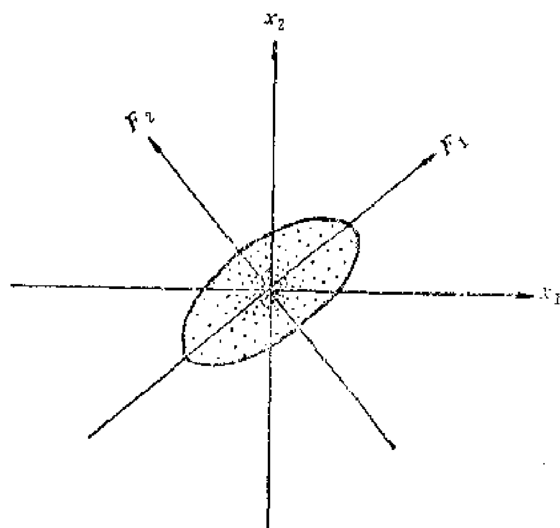


图2-5-1 因子分析的降维

一般情况下，由于因子之间具有比较复杂的相关关系，所以，不宜直接研究这些单个的因子，而是研究由它们的线性组合构成的少数几个综合因子（也称作主因子）。这些综合因子之间相互无关，又能将原有因子所包含的不十分明显的差异尽量多的反映出来。

假设有n个样品，每个样品有m个变量，如果这些变量为$x_1, x_2, \cdots, x_m$，由这m个变量的线性组合构成的综合变量为F_1，F_2，$\cdots$，F_p（$p\leqslant m$）。当$m=2$时，则变量为x_1，x_2，对于二元正态分布变量，n个点在平面上的分布大致为一个椭圆，见图2-5-1。

如果在椭圆的长轴方向选取坐标F_1，在短轴方向选取F_2，这实际上是一次平面坐标变换，变换后显然有如下性质：

（1）二维平面上n个点的方差大部分集中在F_1轴上，而F_2轴上的方差是比较小的。

（2）坐标F_1和F_2之间的相关性很小。

那么，可以称F_1和F_2为原始变量x_1和x_2的综合变量。如果图2-5-1中的椭圆是相当偏平的，则可以只考虑F_1方向上的方差，而忽视F_2方向上的方差。这样，就可以从二维降为一维，而F_1就是x_1和x_2的综合变量。F_1也就可以表示为

$$F_1 = a_{11}x_1 + a_{12}x_2$$

相仿，如果有m个变量x_1，x_2，$\cdots x_m$，则可以将它们表示为p（$p<m$）个综合变量，即

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m \\ \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pm}x_m \end{cases} \tag{2-5-1}$$

此处要求

$$a_{k1}^2 + a_{k2}^2 + \cdots + a_{km}^2 = 1 \qquad (k=1,2,\cdots,p) \tag{2-5-2}$$

系数a_{ij}应由如下原则确定：

（1）F_i与$F_j(i \neq j;\ i=1,\ 2,\ \cdots,\ p)$互相无关。

（2）F_1是$x_1, x_2, \cdots, x_m$的一切线性组合中方差最大的综合变量；F_2是与F_1不相关的x_1，x_2，…，x_m的所有线性组合中方差最大的综合变量；……；F_p是与F_1，$F_2, \cdots F_{p-1}$都不相关的x_1，x_2，…，x_m的所有线性组合中方差最大的综合变量。

这种由线性组合构成的综合变量F_1，F_2，…F_p，称为原变量的第1、第2、…、第p个主因子。其中F_1在总方差中所占的比例最大，其余的主因子F_2，F_3，…，F_p在总方差中所占的比例依次减少。在实际的研究工作中，可以挑选前面的几个最大的主因子作为综合变量，这样就既减少了变量的数目，又抓住了主要矛盾，也简化了变量之间的关系。

由图2-5-1的二维推广到m维空间，从几何学角度看，找主因子的问题就是确定m维空间中椭球体的主轴问题；从代数学角度看，主因子就是x_1，x_2，…x_m的相关矩阵中，p个较大的特征值所对应的特征向量。

一、因子模型

如果已知有n个样品，每个样品观测了m个变量，用x_{ia}表示第a个样品的第i个变量。那么，样品X_a可表为

$$X_a = \begin{pmatrix} x_{1a} \\ x_{2a} \\ \vdots \\ x_{ma} \end{pmatrix} \qquad (a=1,2,\cdots n)$$

而原始资料矩阵为

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} = [x_{ij}]_{m \times n}$$

其中行为变量，列为样品。平均值向量为

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \cdots \\ \bar{x}_m \end{pmatrix}$$

协方差矩阵的估计值是

$$S=\frac{1}{n}(XX'-n\overline{x}\,\overline{x}')=[s_{ij}]_{m\times m}$$

其中 $S_{ij}=\frac{1}{n}\sum_{a=1}^{n}(x_{ia}-\overline{x}_i)(x_{ja}-\overline{x}_j)$

$$(i,\ j=1,\ 2,\ \cdots,\ m)$$

令

$$r_{ij}=\frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

$$(i,\ j=1,\ 2,\ \cdots m)$$

则有 $R=[r_{ij}]_{m\times m}$

R即为相关矩阵。

在实际计算中，首先是将变量进行标准化变换，即

$$x'_{ia}=\frac{x_{ia}-\overline{x}_i}{\sqrt{S_{ii}}}$$

$$(i=1,2,\cdots,m;\ a=1,2,\cdots,n)$$

变换后，x'_{ia}的均值为0，方差为1，这样，协方差矩阵S与相关矩阵R就完全一样了。因此，就可用相关矩阵R进行讨论。R的对角线元素全为1，而m个变量的总方差就是m。

如果变量已经标准化，则原始资料矩阵X与R有如下关系

$$R=XX'$$

记$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$是$R$的特征值，

$$U=[u_1,\ u_2,\ \cdots,\ u_m]$$

是R的特征向量矩阵，则有

$$R=U\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \cdots & \\ & 0 & & \\ & & & \lambda_m\end{pmatrix}U'=XX'$$

两边左乘U'，右乘U就得到

$$U'XX'U=U'U\begin{pmatrix}\lambda_1 & & \\ & \lambda_2 & & 0 \\ & & \cdots & \\ & 0 & & \lambda_m\end{pmatrix}U'U=\begin{pmatrix}\lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \cdots & \\ & 0 & & \lambda_m\end{pmatrix}$$

令$F=U'X$，则有

$$FF'=\begin{pmatrix}\lambda_1 & & & \\ & \lambda_2 & & 0 \\ & & \cdots & \\ & 0 & & \\ & & & \lambda_m\end{pmatrix}$$

那么，F就是相应于原始资料矩阵的主因子矩阵，而F为

$$
\begin{aligned}
F &= [F_1,\ F_2,\ \cdots,\ F_n] \\
&= U'X \\
&= U'[X_1,\ X_2,\ \cdots X_n] \\
F_a &= U'X_a \qquad\qquad (a=1,\ 2,\ \cdots,\ n)
\end{aligned}
$$

即每个F_a就是第a个样品的主成分观测值。

实际研究工作中，为了简化问题，并不一定需要m个主成分，而只用其中一部分就足以代表m个变量的变化情况了，如果选择了其中的p个，可使得

$$\frac{\lambda_1+\lambda_2+\cdots+\lambda_p}{m} \geqslant 85\%\sim 90\%$$

就可以了。此时表明，p个主成分就能够代表全部变量的85%～90%以上的信息，这样一来，m个主成分的矩阵U，将被剖分为两个部分。即

$$U=[\underbrace{u_1,\ u_2,\ \cdots u_p}_{p},\ \underbrace{u_{p+1},\ \cdots,\ u_m}_{m-p}]=[\underset{mp}{U_1},\ \underset{m(m-p)}{U_2}]$$

因而有 $\underset{(p\times n)}{F_{(1)}}=\underset{(p\times m)}{U_1'}\underset{(m\times n)}{X}$

因为 $F=U'X$

所以 $\underset{(m\times n)}{X}=UF=[U_1U_2]\begin{bmatrix}F_{(1)}\\F_{(2)}\end{bmatrix}=\underset{(m\times p)}{U_1}\underset{(p\times n)}{F_{(1)}}+\underset{(m(m-p))}{U_2}\underset{((m-p)n)}{F_{(2)}}$

此式就是资料矩阵X被剖分的结果，可称为因子分析式。其中的$U_1F_{(1)}$是由p个主成分所能解释的部分，而$U_2F_{(2)}$可称为残余部分。

$$
\begin{aligned}
R=XX' &= [U_1F_{(1)}+U_2F_{(2)}][U_1F_{(1)}+U_2F_{(2)}]' \\
&= U_1F_{(1)}F_{(1)}'U_1'+U_1F_{(1)}F_{(2)}'U_2'+U_2F_{(2)}F_{(1)}'U_1' \\
&\quad +U_2F_{(2)}F_{(2)}'U_2'
\end{aligned}
$$

因而$R-U_1F_{(1)}F_{(1)}'U_1'$就是残余的协方差矩阵。如果残余的协方差矩阵很接近对角线矩阵，就表明p个主成分选得比较合适。残余部分可表为

$$\varepsilon=\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_m\end{pmatrix}$$

其中的$\varepsilon_i(i=1,2,\cdots,m)$是相互独立服从正态分布$N(0,\sigma_i^2)$，则有

$$X=U_1F_{(1)}+\varepsilon$$

式中的U_1称为因子载荷系数矩阵，$F_{(1)}$称为主要因子，ε称为特殊因子。

于是，第i个变量可表为如下的线性组合，即$x_i=u_{i1}F_1+u_{i2}F_2+\cdots+u_{ip}F_p+\varepsilon_i$

$$(i=1,2,\cdots,m)$$

习惯上，R型因子模型表示为

$$x_i=a_{i1}F_1+a_{i2}F_2+\cdots\cdots+a_{ip}F_p+a_i\varepsilon_i \qquad (2\text{-}5\text{-}3)$$

$$(i=1,2,\cdots m)$$

类似地，Q型因子模型表示为

$$X_i=a_{i1}F_1+a_{i2}F_2+\cdots+a_{ip}F_p+a_i\varepsilon_i \quad (2\text{-}5\text{-}4)$$

$$(i=1,2,\cdots,n)$$

二、因子模型的含义

因子模型中的F_1，F_2，…，F_p叫作公共因子，它们是在各变量中都共同出现的因子。这些公共因子是相互独立的理论变量，就是前面所述的综合变量（主因子），可以把它们理解为高维空间中互相垂直的P个坐标轴。而ε_1，ε_2，…叫作特殊因子，它们是每个单一变量所特有的因子，各特殊因子之间以及特殊因子与所有公共因子之间都是相互独立的。

a_{ij}叫作因子载荷，是第i个变量在第j个主因子轴上的负荷，如果把x_i看成P维空间中的一个向量，则a_{ij}表示在坐标F_j上的投影。矩阵$A=[a_{ij}]$称为因子载荷矩阵。

1. 因子载荷的统计意义

如果上述原始变量、公共因子、特殊因子均为标准化（平均值为0，方差为1）的变量，则有

$$x_iF_j=a_{i1}F_1F_j+a_{i2}F_2F_j+\cdots+a_{ij}F_jF_j+\cdots+a_{ip}F_pF_j+a_i\varepsilon_jF_j$$

两边取其数学期望E

$$E(x_iF_j)=a_{i1}E(F_1F_j)+a_{i2}E(F_2F_j)+\cdots+a_{ij}E(F_jF_j)+\cdots+a_{ip}E(F_pF_j)$$
$$+a_iE(\varepsilon_iF_j)$$

此时的数学期望即为相关系数r

$$r_{x_iF_j}=a_{i1}r_{F_1F_j}+a_{i2}r_{F_2F_j}+\cdots+a_{ij}r_{F_jF_j}+\cdots+a_{ip}r_{F_pF_j}+a_ir_{\varepsilon_iF_j}$$

由于各公共因子之间具有相互独立的性质，因而有

$$r_{F_1F_j}=r_{F_2F_j}=\cdots=r_{F_pF_j}=r_{\varepsilon_iF_j}=0$$

只有 $r_{F_jF_j}=1$

所以 $r_{x_iF_j}=a_{ij}$

可见，因子载荷反映了变量（样品）与主因子之间的关系，a_{ij}就是第i个变量与第j个公共因子的相关系数。

2. 变量共同度的统计意义

因子载荷矩阵A中各行元素的平方和

$$h_i^2=\sum_{j=1}^{p}a_{ij}^2 \quad (2\text{-}5\text{-}5)$$

$$(i=1,\ 2,\ \cdots,\ m)$$

称为变量x_i的共同度。为了研究它的统计意义，可对R型因子分析的变量x_i计算其方差，由于标准化变量的方差等于1，所以有

$$Dx_i=a_{i1}^2DF_1+a_{i2}^2DF_2+\cdots+a_{ip}^2DF_p+a_i^2D\varepsilon_i$$
$$=a_{i1}^2+a_{i2}^2+\cdots+a_{ip}^2+a_i^2$$

$$= \sum_{j=1}^{p} a_{ij}^2 + a_i^2$$

$$= h_i^2 + a_i^2 = 1 \qquad (2\text{-}5\text{-}6)$$

可见，变量x_i的方差由两部分所组成，第一部分为共同度h_i^2，它是全部公共因子对变量x_i的总方差所作的贡献；第二部分是特定变量所产生的方差，称作特殊因子方差，它仅与变量x_i本身的变化有关，也就是说，它是使变量x_i的方差为1的补充值。

3. 公共因子F_j的方差贡献的统计意义

因子载荷矩阵中各列元素的平方和

$$s_j = \sum_{i=1}^{m} a_{ij}^2$$

$$(j=1,2,\cdots,P)$$

的统计意义，与变量x_i的共同度h_i^2恰好相反，h_i^2是诸公共因子对同一变量x_i所提供的方差的总和。而s_j则是同一公共因子F_j对诸变量所提供的方差的总和，它是衡量各个公共因子相对重要性的标准。

三、因子模型的几何意义

为了深入理解因子模型，有必要从几何学角度进行讨论。

在因子分析中，可以把互不相关的，即两两之间的相关系数（夹角余弦）为0，而各自方差为1的p个公共因子和m个特殊因子想象成（$p+m$）个相互垂直的（即两两之间夹角余弦为0）的单位向量，以它们为坐标轴，就构成了（$p+m$）维空间的一个直角坐标系，这些坐标轴可称为因子轴，称这个（$p+m$）维空间为因子空间。因而，变量x_i可用因子空间中的向量P_i表示，向量P_i的模等于1。即

$$|P_i| = \sqrt{a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 + a_i^2} = 1$$

此时，P_i与各因子轴的夹角余弦就等于x_i与各因子的相关系数，即

$$r_{P_iF_j} = \cos(P_iF_j) = |P_i|\cos(P_iF_j) = a_{ij} = r_{x_iF_j} \qquad (2\text{-}5\text{-}7)$$

另外，因子空间中变量x_i、x_j的向量P_i、P_j的夹角余弦正好是两个变量x_i、x_j的相关系数，即

$$r_{P_iP_j} = \cos(P_iP_j) = \frac{P_iP_j}{|P_i||P_j|} = P_iP_j$$

$$= a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{ip}a_{jp} = r_{x_ix_j} \qquad (2\text{-}5\text{-}8)$$

第二节　主因子解

因子分析所要讨论的基本问题是用变量之间的相关系数决定因子载荷。因此，必须建立二者之间的关系。m个变量x_1，x_2，$\cdots x_m$的相关系数矩阵为

$$R=\begin{pmatrix}1 & r_{12} & \cdots & r_{1m}\\ r_{21} & 1 & \cdots & r_{2m}\\ \cdots & \cdots & \cdots & \cdots\\ r_{m1} & r_{m2} & \cdots & 1\end{pmatrix}$$

前已述及，对于已标准化的变量有

$$r_{ij}=a_{i1}a_{j1}+a_{i2}a_{j2}+\cdots+a_{ip}a_{jp}=\sum_{K=1}^{p}a_{iK}a_{jK} \qquad (i\neq j)$$

当$i=j$时有

$$r_{jj}=a_{j1}^2+a_{j2}^2+\cdots+a_{jp}^2=h_j^2+a_j^2$$

因此

$$R=\begin{pmatrix}1 & r_{12} & \cdots & r_{1m}\\ r_{21} & 1 & \cdots & r_{2m}\\ \cdots & \cdots & \cdots & \cdots\\ r_{m1} & r_{m2} & \cdots & 1\end{pmatrix}$$

$$=\begin{pmatrix}h_1^2+a_1^2 & \sum_{K=1}^{p}a_{1K}a_{2K} & \cdots & \sum_{K=1}^{p}a_{1K}a_{mK}\\ \sum_{K=1}^{p}a_{2K}a_{1K} & h_2^2+a_2^2 & \cdots & \sum_{K=1}^{p}a_{2K}a_{mK}\\ \cdots & \cdots & \cdots & \cdots\\ \sum_{K=1}^{p}a_{mK}a_{1K} & \sum_{K=1}^{p}a_{mK}a_{2K} & \cdots & h_m^2+a_m^2\end{pmatrix}$$

$$=\begin{pmatrix}h_1^2 & \sum_{K=1}^{p}a_{1K}a_{2K} & \cdots & \sum_{K=1}^{p}a_{1K}a_{mK}\\ \sum_{K=1}^{p}a_{2K}a_{1K} & h_2^2 & \cdots & \sum_{K=1}^{p}a_{2K}a_{mK}\\ \cdots & \cdots & \cdots & \cdots\\ \sum_{K=1}^{p}a_{mK}a_{1K} & \sum_{K=1}^{p}a_{mK}a_{2K} & \cdots & h_m^2\end{pmatrix}+\begin{pmatrix}a_1^2 & & & 0\\ & a_2^2 & & \\ & & \cdots & \\ 0 & & & a_m^2\end{pmatrix}$$

$$=\begin{pmatrix}a_{11} & a_{12} & \cdots & a_{1p}\\ a_{21} & a_{22} & \cdots & a_{2p}\\ \cdots & \cdots & \cdots & \cdots\\ a_{m1} & a_{m2} & \cdots & a_{mp}\end{pmatrix}\begin{pmatrix}a_{11} & a_{21} & \cdots & a_{m1}\\ a_{12} & a_{22} & \cdots & a_{m2}\\ \cdots & \cdots & \cdots & \cdots\\ a_{1p} & a_{2p} & \cdots & a_{mp}\end{pmatrix}+\begin{pmatrix}a_1^2 & & & \\ & a_2^2 & & 0\\ & & \cdots & \\ 0 & & & \\ & & & a_m^2\end{pmatrix}$$

$$=AA'+aa'$$

如果不考虑特殊因子，可取

$$R^* = R - aa' = AA' \qquad (2\text{-}5\text{-}9)$$

这里称R^*为约相关矩阵，它与R的区别仅在于对角线元素，R^*的对角线元素依次为变量共同度h_i^2。可见，因子分析就是在已知约相关矩阵R^*的条件下，求解因子载荷矩阵A，使得$R^*=AA'$这便是因子分析的基本出发点。

如果A是R^*的解，而C是一个任意的（PXP）阶的正交矩阵，则有

$$(AC)(AC)' = (AC)(A'C') = A(CC')A' = AA'$$

这一情况，说明（2-5-9）式具有多解性。因此，需要首先讨论一下主因子解。

假定$aa'=0$，此时，可由相关矩阵$R=AA'$求出主因子解。矩阵R的元素为

$$r_{ij} = \sum_{K=1}^{P} a_{iK}a_{jK} \qquad (2\text{-}5\text{-}10)$$

$$(i, j = 1, 2, \cdots, m)$$

所谓主因子解就是根据变量之间的相关矩阵中选出第一个主因子F_1，并使其在各变量的公共因子方差中所占的方差贡献为最大；然后消去这个因子的影响，再从剩余因子的相关矩阵中，选出与F_1不相关的因子F_2，使其在各个变量的剩余公共因子方差中贡献为最大；这样继续选择F_3，F_4……，直到各个变量的公共因子方差被分解完毕为止。

挑选的第一个主因子F_1，必须使它的方差贡献$S_1=\sum_{i=1}^{m} a_{i1}^2$在（2-5-10）式条件下为最大。这是个条件极值问题，常用的计算方法是拉格朗日乘数法。令

$$2T = S_1 - \sum_{i,j=1}^{m} \mu_{ij} r_{ij} = S_1 - \sum_{i,j=1}^{m} \sum_{K=1}^{P} \mu_{ij} a_{iK} a_{jK}$$

式中$\mu_{ij}=\mu_{ji}$为拉格朗日乘数，T为一新函数，求T对每一个变量a_{i1}的偏导数，且令其等于0，即

$$\frac{\partial T}{\partial a_{i1}} = a_{i1} - \sum_{j=1}^{m} \mu_{ij} a_{j1} = 0$$

同样，也可以求T对其余每个变量a_{iK}（$K \neq 1$）的偏导数，且令其等于0，即

$$\frac{\partial T}{\partial a_{iK}} = -\sum_{j=1}^{m} \mu_{ij} a_{jK} = 0 \qquad (K \neq 1)$$

将上面两个方程式结合起来可写成

$$\frac{\partial T}{\partial a_{iK}} = \delta_{1K} a_{i1} - \sum_{j=1}^{m} \mu_{ij} a_{jK} = 0 \qquad (2\text{-}5\text{-}11)$$

$$(K = 1, 2, \cdots, P)$$

上式中的 $\delta_{1K} = \begin{cases} 1 & (K=1) \\ 0 & (K \neq 1) \end{cases}$

用a_{i1}乘（2-5-11）式两边，并对i求和，得到

$$\delta_{1K}\sum_{i=1}^{m}a_{i1}^{2}-\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_{ij}a_{i1}a_{jK}=0$$

由（2-5-11）式有

$$\sum_{i=1}^{m}\mu_{ij}a_{i1}=a_{j1}$$

从而上式变为

$$\delta_{1K}S_{1}-\sum_{j=1}^{m}a_{j1}a_{jK}=0 \qquad (2\text{-}5\text{-}12)$$

再用a_{iK}乘此式两边并对K求和，则有

$$a_{i1}S_{1}-\sum_{j=1}^{m}a_{j1}\left(\sum_{K=1}^{p}a_{iK}a_{jK}\right)=0$$

$$(i=1,\ 2,\ \cdots,\ m)$$

应用（2-5-10）式有

$$\sum_{j=1}^{m}r_{ij}a_{j1}-s_{1}a_{i1}=0$$

$$(i=1,\ 2,\ \cdots,\ m)$$

写成矩阵形式就是

$$\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{pmatrix}\begin{pmatrix} a_{11} \\ a_{21} \\ \cdots \\ a_{m1} \end{pmatrix}-s_{1}\begin{pmatrix} a_{11} \\ a_{21} \\ \cdots \\ a_{m1} \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix} \qquad (2\text{-}5\text{-}13)$$

记 $a_{1}=[a_{11},\ a_{21},\ \cdots,\ a_{m1}]'$，$0=[0,\ 0,\ \cdots,\ 0]'$

用I表示（$P\times P$）阶单位矩阵，则（2-5-13）式可简化为

$$(R-s_{1}I)\,\mathbf{a}_{1}=0$$

或者

$$R\mathbf{a}_{1}=s_{1}\mathbf{a}_{1}$$

其中R为相关矩阵，$a_{i1}(i=1,\ 2,\cdots,m)$为因子载荷，根据主因子要求，a_{11}，a_{21}，$\cdots$，a_{m1}不应同时为0，因此，线性方程组（2-5-13）式的系数行列式必须为0，亦即s_{1}应满足条件

$$|R-s_{1}I|=0$$

这就是相关矩阵R的特征方程。由于要求s_{1}为最大，所以s_{1}应等于R的最大特征值λ_{1}，即

$$s_{1}=\sum_{i=1}^{m}a_{i1}^{2}=\lambda_{1} \qquad (2\text{-}5\text{-}14)$$

选出第一个公共因子F_{1}之后，假如各变量的公共因子方差未被分解完毕，就要继续选择第二个公共因子F_{2}，它与F_{1}互不相关，而且其方差贡献s_{2}要在条件

$$R_{1}=R-\mathbf{a}_{1}\mathbf{a}_{1}'$$

或者在

$$r_{ij}^{(1)}=r_{ij}-a_{i1}a_{j1}=\sum_{K=2}^{P}a_{iK}a_{jK}$$

$$(i,\ j=1,\ 2,\ \cdots,\ m)$$

之下为最大。R_1是从R中扣除F_1的影响之后的剩余相关矩阵，$r_{ij}^{(1)}$为R_1中第i行第j列的元素。因此，可用类似于选F_1的办法来选F_2。事实上，矩阵R_1与R的特征向量相同，而且除了对应于R_1中的特征向量a_1的特征值为0而不是特征值λ_1外，矩阵R_1与矩阵R的所有特征值都相同。

假如$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0$为$R$矩阵的$P$个特征值，$a_1$，$a_2$，……$a_P$表示与其对应的规格化的特征向量。因为对于任何的$K=1,\ 2,\ \cdots,\ P$都有

$$R\boldsymbol{a}_K=\lambda_K\boldsymbol{a}_K$$

而

$$R_1a_K=(R-a_1a_1')a_K=Ra_K-a_1a_1'a_K=\lambda_Ka_K-a_1a_1'a_K \qquad (2\text{-}5\text{-}15)$$

这里分别对$K=1$与$K\neq1$两种情况进行讨论：

（1）当$K=1$时，由（2-5-12）、(2-5-14)式有$a_1'a_1=\lambda_1$，因而(2-5-15)式可化为$R_1a_1=0$，即对应于R中的最大特征值的特征向量在R_1中为0。

（2）当$K\neq1$时，由于$a_1'a_1=0$，（2-5-15）式变为$R_1a_K=\lambda_Ka_K$。

通过上述两种情况的讨论可见，R_1与R的特征向量相同；除了特征值λ_1外，其他的特征值也相同。因此，R的次大特征值λ_2就是R_1的最大特征值。那么，取R的次大特征值λ_2与其对应的特征向量a_2，则第二个公共因子F_2就确定了。同样，其后的各个公共主因子F_3，F_4…，F_P都可以由相关矩阵R求其特征值解而确定。可见，通过对R进行特征值分析就可以将全部的公共因子找出。

上述讨论都是以R型因子分析为例进行的，对于Q型因子分析也有相类似的结果。

第三节　方差最大正交旋转

通过因子分析可以找出主要的公共因子，但是，更重要的是要明确每个主因子的实际地质意义。为此，还需要对因子载荷施行旋转使其结构化，使每个因子载荷的平方向1或0两极分化，其中的第j个主因子的代表性变量在F_j因子轴上的载荷系数等于或趋近于1，而在其他因子轴上的载荷系数等于或趋近于0。这样便容易对每个主因子进行地质解释。此种作法，从数学上来说就是对矩阵A进行正交变换，亦即将各因子轴在它们所确定的空间中作一正交旋转。

目前，对因子载荷的旋转方法有许多种，但其中最常用方法是方差最大正交旋转，这种旋转方法是使因子载荷矩阵中的各因子载荷值的方差达到最大。对于R型因子载荷矩阵$A=[a_{ij}]_{m\times p}$，对因子F_j的简化，可由因子载荷值的平方方差来表示，即

$$V_j=\frac{1}{m}\sum_{i=1}^{m}(b_{ij}^2)^2-\left(\frac{1}{m}\sum_{i=1}^{m}b_{ij}^2\right)^2$$

$$=\left[m\sum_{i=1}^{m}(b_{ij}^2)^2-\left(\sum_{i=1}^{m}b_{ij}^2\right)^2\right]\Big/m^2$$

式中的b_{ij}是经过正交旋转后所得因子载荷矩阵B的元素。使用载荷值的平方是为了避免出现负值。

如果使V_j为最大，亦即使第j个因子得到最大的简化，此时，它在因子空间中F_j的载荷系数趋于1，而在其它因子轴上的载荷系数趋于0。那么，对于整个因子矩阵 $A=[a_{ij}]_{m\times p}$的简化则可由所有因子载荷的平方方差之和作为衡量标准，即使

$$V=\sum_{j=1}^{p}V_j=\sum_{j=1}^{p}\left[m\sum_{i=1}^{m}(b_{ij}^2)^2-\left(\sum_{i=1}^{m}b_{ij}^2\right)^2\right]\Big/m^2 \qquad (2\text{-}5\text{-}16)$$

达到最大。考虑到各个变量$x_i(i=1,2,\cdots,m)$的共同度之间的差异，需要用（b_{ij}/h_{ij}^2）来代替（2-5-16）式中的b_{ij}^2，这实际上是需要求得经过旋转后的b_{ij}，使其

$$V=\sum_{j=1}^{p}\left[m\sum_{i=1}^{m}(b_{ij}^2/h_{ij}^2)^2-\left(\sum_{i=1}^{m}(b_{ij}^2/h_{ij}^2)\right)^2\right]\Big/m^2 \qquad (2\text{-}5\text{-}17)$$

达到最大。

对因子载荷矩阵$A=[a_{ij}]_{m\times p}$进行正交旋转，相当于对所有因子面$F_gF_q(g=1,2,\cdots m-1;\ q=g+1,\ g+2,\ \cdots,\ m)$正交旋转一个角度$\varphi_{gq}$，每次的旋转角$\varphi$必须满足使（2-5-17）式中的$V$达到最大值。为此，可选择如下的正交变换

$$T_{gq}=\begin{pmatrix}1 & & & & & \\ & \ddots & & & & \\ & & \cos\varphi & & -\sin\varphi & \\ & & & \ddots & & \\ & & \sin\varphi & & \cos\varphi & \\ & & & & & \ddots \\ & & & & & & 1\end{pmatrix}_{m\times m}\begin{matrix} \\ \\ g \\ \\ q \\ \\ \\ \end{matrix}$$

$$\qquad\qquad g \qquad\qquad q$$

T_{gq}中凡没有标明的元素均为0，A经过变换后，相当于将因子平面F_gF_q旋转一个角度φ，得到矩阵

$$B=AT_{gq}=[b_{ij}]_{m\times p}$$

B中的元素分别为

$$\begin{aligned}&b_{ig}=a_{ig}\cos\varphi+a_{iq}\sin\varphi\\&b_{iq}=-a_{ig}\sin\varphi+a_{iq}\cos\varphi \qquad\qquad (2\text{-}5\text{-}18)\\&b_{iK}=a_{iK} \qquad\qquad (K\neq g,\ q;\ i=1,\ 2,\ \cdots,\ m)\end{aligned}$$

如果有P个主因子，则必须对A中所有p列全部配对旋转，总共旋转

$$C_p^2=\frac{p(p-1)}{2}$$

次，全部旋转完毕算一个循环，此时，得到载荷矩阵

$$B_1=AT_{12}\cdots T_{1p}\cdots T_{(p-1)p}=A\prod_{g=1}^{p-1}\prod_{q=g+1}^{p}T_{gq}=AC_1$$

上式中记$C_1=\prod_{g=1}^{p-1}\prod_{q=g+1}^{p}T_{gq}$，$B_1$为对$A$进行正交变换$C_1$而得。

经过第一个循环后，可按（2-5-17）式计算V_1。在第一个循环基础上，从B_1出发进行第二个旋转循环，旋转完成后得到B_2，即

$$B_2=B_1\prod_{g=1}^{p-1}\prod_{q=g+1}^{p}T_{gq}=B_1C_2=AC_1C_2$$

由B_2可计算出V_2。

如此不断地重复这个步骤，就可以得到V的一个非降有界序列

$$V_1\leqslant V_2\leqslant\cdots$$

众所周知，由于因子载荷的绝对值不大于1，所以这个序列是有上界的，必然收敛于某一极限V_{max}，V_{max}为V的极大值，因此，只要循环次数k充分大，就必然有

$$|V_k-V|<\varepsilon$$

ε为所要求的计算精度。如果循环次数k与（$k+1$）都充分大时，也必然有

$$|V_k-V_{k+1}|<\varepsilon$$

最后得 $B_k=A\prod_{i=1}^{k}C_i=AC$

B_k就是旋转后的因子载荷矩阵。

前已述及，在任何一次变换T_{gq}中，必须使方差达到极大，为此，应按如下步骤确定旋转角度：

（1）将（2-5-18）式代入（2-5-17）式；

（2）将（2-5-17）式对φ求一阶导数并令其为0，可解得

$$\mathrm{tg}4\varphi=\frac{D-2AB/m}{C-(A^2-B^2)/m}=\frac{E}{F}\qquad（2\text{-}5\text{-}19）$$

上式中的m为变量个数。令

$$T_i=(a_{ig}/h_i)^2-(a_{iq}/h_i)^2$$
$$H_i=2(a_{ig}/h_i)(a_{iq}/h_i)$$

则

$$A=\sum_{i=1}^{m}T_i$$

$$B=\sum_{i=1}^{m}H_i$$

$$C=\sum_{i=1}^{m}(T_i^2-H_i^2)$$

$$D=2\sum_{i=1}^{m}T_iH_i$$

（3）将（2-5-17）式展开，并将包含φ的项合并简化，最后就只剩下包含$\sin4\varphi$和$\sin^22\varphi$的项，使（2-5-17）式成为以$\frac{\pi}{2}$为周期的函数。因而，（2-5-19）式中的4φ只要在$\frac{\pi}{2}$的范围内考虑就行。通常在$-\frac{\pi}{4}\sim\frac{\pi}{4}$之间考虑，同时由（2-5-17）式对$\varphi$的二阶导数应小于0，

可得

$$\frac{1}{E}\sin 4\varphi > 0$$

所以，φ的符号可根据E的符号确定，它应与E同号，所以，可按分子E及分母F的正负号来确定4φ应在哪一象限中。

以上以R型因子分析为例，对方差最大正交旋转问题进行了讨论。对于Q型因子分析，载荷矩阵的最大正交旋转情况相仿，Q型因子分析$A=[a_{ij}]_{n\times p}$，所以，只要将以上公式中的m换成n即可。

第四节　因子得分

因子分析是将变量（或样品）表示为公共因子的线性组合。当然，反过来也可以将公共因子表示为变量（或样品）的线性组合。即

$$F_j = C_{j1}x_1 + C_{j2}x_2 + \cdots + C_{jm}x_m \qquad (2\text{-}5\text{-}20)$$

$$(j=1,\ 2,\ \cdots,\ p)$$

F_j称为因子得分函数，它是一个综合指标，可以把原始变量中的关于公共因子F_j的信息集中起来。如果取$m=2$，即取p个主因子中的两个因子，则可将每一样品点的m个变量代入（2-5-20）式，计算出因子得分F_i及F_j（$i\neq j$；$i,\ j=1,\ 2,\ \cdots,\ p$），若在二维平面上点图，就能对样品进行分类或者进行地质解释。

如果（2-5-20）式中方程个数与未知量F_j的个数相等，则容易解出F_j（$j=1,\ 2,\ \cdots,\ p$），但往往是$p<m$，即方程个数少于变量个数，这样会出现矛盾方程组，而不能根据变量取值来准确计算因子得分。因此，只能在最小二乘法意义下对因子得分进行估计。为此，必须建立F_j（$j=1,\ 2,\ \cdots,\ p$）关于变量$x_1,\ x_2,\ \cdots,\ x_m$的回归方程

$$\hat{F}_j = b_{j0} + b_{j1}x_1 + b_{j2}x_2 + \cdots + b_{jm}x_m \qquad (2\text{-}5\text{-}21)$$

$$(j=1,\ 2,\ \cdots,\ p)$$

由于变量和主因子都已经过标准化，因而

$$b_{j0} = \overline{F}_j - b_{j1}\bar{x}_1 - b_{j2}\bar{x}_2 - \cdots - b_{jm}\bar{x}_m = 0$$

所以（2-5-21）式中不会有常数项，于是

$$\hat{F}_j = [b_{j1},\ b_{j2},\ \cdots,\ b_{jm}][x_1,\ x_2,\ \cdots,\ x_m]' = b_j'X$$

$$(j=1,\ 2,\ \cdots,\ p)$$

对于标准化变量和因子来说，有

$$\hat{F}_j = (S^{-1}B_j)'X = (R^{-1}B_j)'X = B_j'R^{-1}X$$

$$(j=1,\ 2,\ \cdots,\ p)$$

其中　$B_j' = [r_{x_1F_j},\ r_{x_2F_j},\ \cdots,\ r_{x_mF_j}] = [a_{1j},\ a_{2j},\ \cdots,\ a_{mj}]$

最后有

$$\hat{F}=\begin{pmatrix}\hat{F}_1\\ \hat{F}_2\\ \vdots\\ \hat{F}_p\end{pmatrix}=\begin{pmatrix}B'_1\\ B'_2\\ \vdots\\ B'_m\end{pmatrix}R^{-1}X=\begin{pmatrix}a_{11} & a_{21} & \cdots & a_{m1}\\ a_{21} & a_{22} & \cdots & a_{m2}\\ \cdots & \cdots & \cdots & \cdots\\ a_{1p} & a_{2p} & \cdots & a_{mp}\end{pmatrix}R^{-1}X$$

$$=A'R^{-1}X \qquad (2\text{-}5\text{-}22)$$

式中

$$\hat{F}=[\hat{F}'_1,\ \hat{F}_2,\ \cdots,\ \hat{F}_p]'$$

$$X=[x_1,\ x_2,\ \cdots,\ x_m]'$$

$$R=\begin{pmatrix}r_{x_1x_1} & r_{x_1x_2} & \cdots & r_{x_1x_m}\\ r_{x_2x_1} & r_{x_2x_2} & \cdots & r_{x_1x_m}\\ \cdots & \cdots & \cdots & \cdots\\ r_{x_mx_1} & r_{x_mx_2} & \cdots & r_{x_nx_m}\end{pmatrix}$$

R为原变量的相关系数矩阵；$A'=[a_{ij}]_{p\times m}$。

当因子之间为正交时，所得到的矩阵就是旋转后的因子载荷矩阵的转置矩阵。

第五节　算　　例

本算例是对周口盆地白垩系、下第三系生油岩抽提物和油样的与甾烷有关的地球化学分析数据，进行R型因子分析，研究其母质类型和成熟度及运移效应。

选取5个分析项目作为变量，其意义如下：

x_1——甾烷异构化参数（$\alpha\alpha\beta$R/$\alpha\alpha\alpha$R），可作为生油层的成熟度标志：

$x_1<0.25$　未成熟

$0.25\leqslant x_1<0.42$　低成熟

$0.42\leqslant x_1$　成熟

x_2——C_{29}规则甾烷（$\alpha\beta\beta$R/$\alpha\alpha\alpha$R），为运移效应标志，可判断烃类运移效应和运移距离的相对远近，其值越大，说明运移效应越明显。

x_3——C_{27}规则甾烷分布（%）。

x_4——C_{28}规则甾烷分布（%）。

x_5——C_{29}规则甾烷分布（%）。

x_3、x_4、x_5为母质类型分布指数，富C_{27}而贫C_{29}，则生油母源以低等水生生物为主，母质类型较好；富C_{29}而贫C_{27}，则母源中高等植物较多，母质类型较差。

取8口井的33个样品，除12、19、25号样品为油样外，其余样品均为生油岩的抽提物。原始数据见表2-5-1。

表2-5-1　8口井33个样品的化验分析数据

样品序号	x_1	x_2	x_3	x_4	x_5
1	0.424	0.550	24.860	24.620	50.520
2	0.411	0.598	26.600	24.530	48.460
3	0.394	0.460	26.800	23.650	49.550
4	0.414	0.562	25.110	24.670	48.460
5	0.425	0.833	26.220	29.500	44.200
6	0.433	0.911	18.480	31.630	49.890
7	0.438	0.420	24.770	22.520	52.710
8	0.422	0.668	23.790	30.320	45.880
9	0.448	0.576	19.750	24.550	65.700
10	0.497	0.647	22.090	29.880	48.030
11	0.352	0.470	25.390	23.770	50.840
12	0.420	0.945	22.140	21.220	56.640
13	0.292	0.460	26.050	26.120	47.830
14	0.395	0.610	28.440	23.210	48.350
15	0.339	0.450	30.980	20.170	48.450
16	0.397	0.545	24.960	25.080	49.970
17	0.398	0.541	28.520	16.500	54.980
18	0.405	0.661	27.760	24.220	48.020
19	0.335	0.595	28.850	18.000	53.160
20	0.433	0.441	22.840	24.270	52.900
21	0.287	0.471	28.350	25.270	46.390
22	0.126	0.201	28.890	25.400	45.710
23	0.227	0.329	23.360	22.140	54.500
24	0.139	0.269	36.010	23.880	40.120
25	0.229	0.159	39.730	24.440	35.830
26	0.153	0.108	34.640	16.130	49.230
27	0.330	0.455	27.390	24.410	48.200
28	0.131	0.164	41.240	21.460	37.300
29	0.482	1.854	33.130	21.600	45.270
30	0.145	0.184	21.260	20.100	52.630
31	0.106	0.134	33.750	16.990	49.270
32	0.265	0.280	38.180	23.380	38.450
33	0.299	0.526	38.180	20.970	41.210

经计算得到方差极大正交旋转因子载荷矩阵，见表2-5-2。变量的因子载荷图见图2-5-2。

由表2-5-2可以看出，第一个主因子的代表性变量为x_2；第二个主因子的代表性变量为x_5；第三个主因子的代表性变量为x_3。

为了对生油岩进行分类和进行进一步的解释，可对R型因子分析的结果计算前三个主因子的样品的因子得分，见表2-5-3。

表2-5-2　方差极大正交旋转因子载荷矩阵

变量序号	主因子		
	F_1	F_2	F_3
x_1	0.8404	0.4188	−0.3440
x_2	0.9966	0.0481	−0.0664
x_3	−0.1905	−0.8991	0.3942
x_4	0.1876	0.0047	−0.9822
x_5	0.1371	0.9598	0.2448

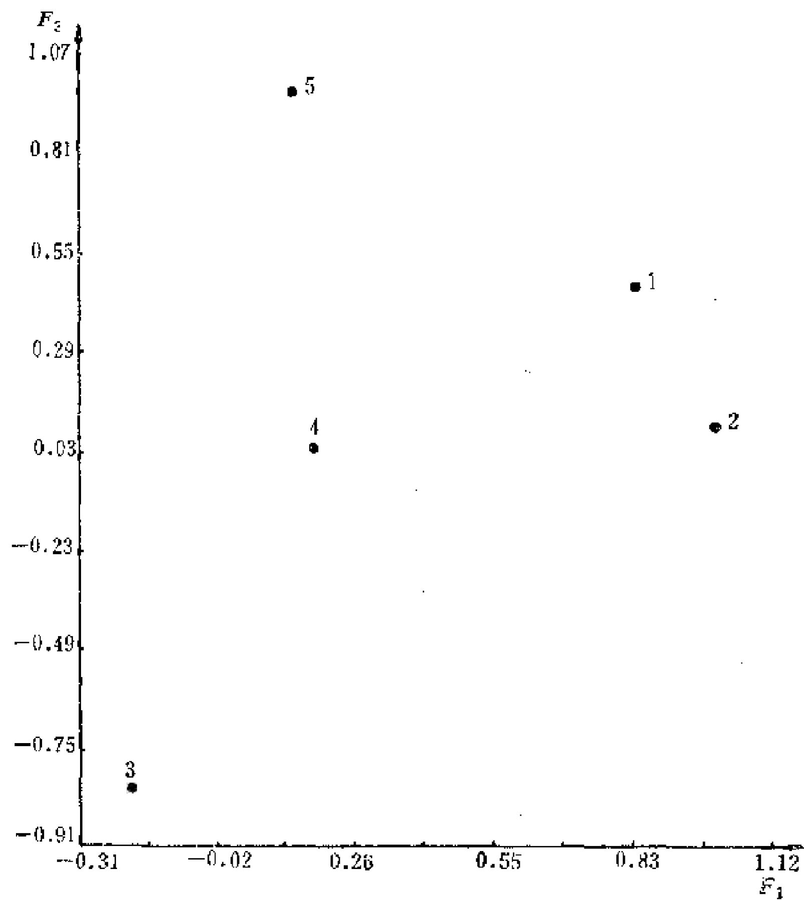


图2-5-2　因子载荷图

表2-5-3　33个样品的因子得分

样品序号	因子得分		
	F_1	F_2	F_3
1	−22.6459	−20.3899	−19.5561
2	−22.2182	−22.2682	−16.4810
3	−22.3671	−18.7061	−29.4380
4	−22.2553	−19.5313	−21.3385
5	−22.4148	−22.3476	−9.1772
6	−24.5376	−20.6710	−19.0977
7	−22.8939	−19.4432	−27.2622
8	−23.2186	−19.2829	−25.9169
9	−24.2134	−8.5119	16.7124
10	−23.6462	−13.8934	10.4893
11	−22.7749	−14.0234	17.5181
12	−23.2164	−9.3178	20.6489
13	−22.7040	−16.0067	14.7607
14	−21.7975	−16.9639	18.1529
15	−21.0407	−18.0577	21.4332
16	−22.8539	−14.2849	16.0148
17	−21.6407	−13.3781	25.9396
18	−21.9712	−16.8084	17.0006
19	−21.5834	−14.5615	24.2147
20	−23.4650	−11.5833	16.9616
21	−22.0465	−17.9704	15.7419
22	−22.1850	−18.5700	15.5296
23	−23.4288	−10.9705	19.4187
24	−20.1347	−26.2637	17.2079
25	−19.1525	−29.4516	16.4532
26	−20.3899	−19.5561	26.1934
27	−22.2682	−16.4810	16.7421
28	−18.7061	−29.4380	19.9106
29	−19.5313	−21.3385	20.2667
30	−22.3476	−9.1772	19.0727
31	−20.6710	−19.0977	25.2250
32	−19.4432	−27.2622	17.7411
33	−19.2829	−26.9169	20.7627

根据计算结果，三个主因子的地质意义如下：

F_1为运移效应及成熟度；

F_2为母质类型及运移效应；

F_3为生油岩的母质类型。

由样品的因子得分计算结果，将三个主因子的代表性样品列于表2-5-4中。在该表中也列出了样品的5个变量值以及样品的地质类型。

表2-5-4　主因子的代表性样品及地质类型

主因子	代表性样品号	因子得分	x_1	x_2	x_3	x_4	x_5	地　质　类　型
F_1	28	−28.7061	0.131	0.164	41.240	21.460	37.300	母质好，未成熟，未运移
F_2	9	−8.5119	0.448	0.576	19.750	24.550	55.700	母质差，已成熟，经过运移
F_3	26	26.1934	0.153	0.108	36.640	16.130	49.230	母质中等，未成熟，未运移

33个样品中的3个油样，其变量值及地质类型列于表2-5-5。

表2-5-5　三个油样的地质类型

样品序号	x_1	x_2	x_3	x_4	x_5	地　质　类　型
12	0.420	0.945	22.140	21.220	56.640	母质差，已成熟，运移明显
19	0.335	0.595	28.860	18.000	53.160	母质中等，低成熟，经过运移
25	0.229	0.159	39.730	24.440	35.830	母质好，未成熟，未经运移

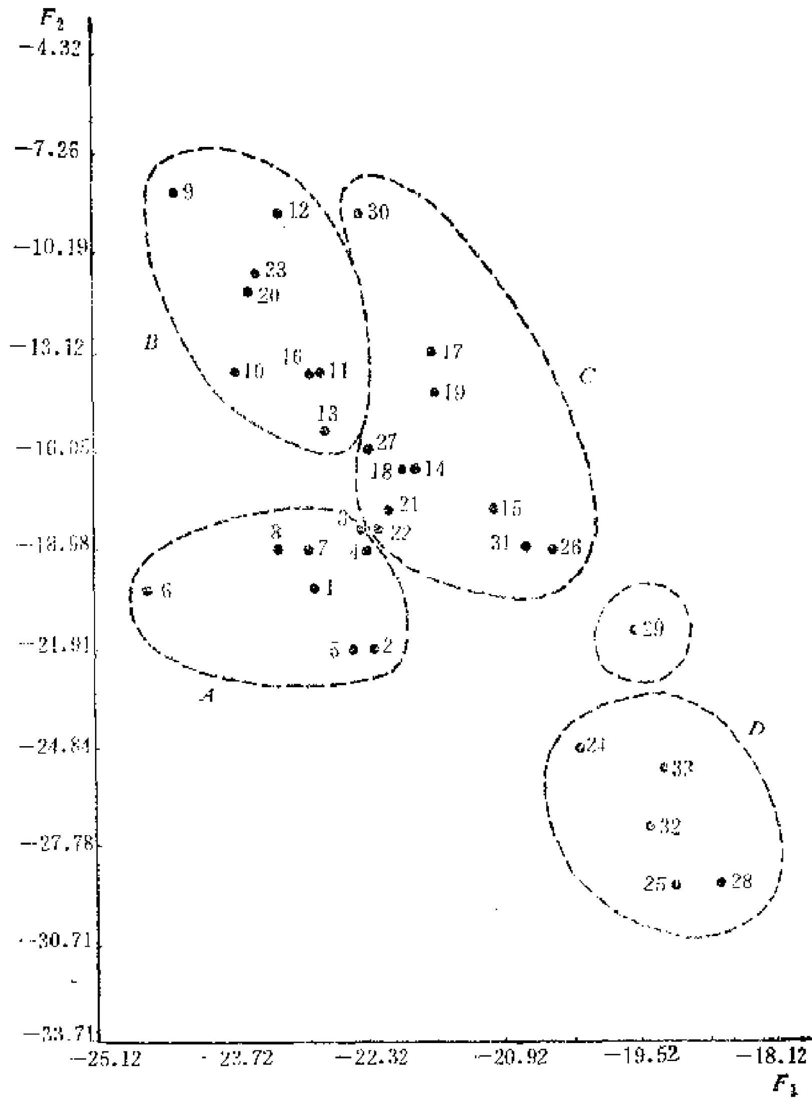


图2-5-3　因子得分图

在$F_1 \sim F_2$主因子样品得分图上，除29号样品外，其余样品可划分为四个样品集团，除A集团外，B、C、D集团中各包含一个已知地质类型的代表性样品，见图2-5-3。

根据各样品集团变量观测值的地质意义，将各集团的地质类型列于表2-5-6中。

表2-5-6 各样品集团的地质类型

集团号	集团中的样品号	代表样品号	地质类型
A	1, 2, 3, 4, 5, 6, 7, 8,		母质差到中等，成熟到低成熟，经过明显运移
B	9, 10, 11, 12, 13, 16, 20, 23	9, 12	母质差，低成熟，少数成熟，经过明显运移
C	14, 15, 17, 18, 19, 21, 22, 26, 27, 30, 31	19, 26	母质中等，低成熟，少数未成熟，经过运移
D	24, 25, 28, 32, 33	25, 28	母质好，未成熟，少数低成熟，未经运移
	29		母质较好，已成熟，经远距离运移

通过上述研究，对周口盆地的生油条件和运移状况可以得出如下认识：

（1）白垩系生油岩的母质类型较差，为成熟到低成熟。其中，位于生油凹陷东部的母质类型以差为主，而靠近凹陷中部的母质类型相对较好，说明盆地沉积中心生油条件较好。

（2）白垩系生油岩已达到成熟或低成熟，凹陷中部成熟程度相对较高。

（3）白垩系生油岩生成的烃类普遍经历了运移。在适当条件下，有可能形成油气藏。

（4）第三系生油岩的母质类型较好，但未成熟，且未经过运移。部分层位样品表现为母质中等，低成熟，并显示出一定的运移效应，反映这些层位曾受白垩系生油岩生成的烃类浸染。

第六章　对应分析

对应分析是在因子分析基础上发展起来的一种多元统计分析方法，它是把R型与Q型因子分析结合起来，对变量与样品统一进行分析研究的方法，因而更有利于进行地质解释。

前已述及，因子分析可以用少数的几个公共因子去提取研究对象的绝大部分信息，因而，既可减少因子的数目，又能把握住研究对象之间的相互关系。而且通过研究公共因子的特征，比较容易揭示研究对象在成因上或空间上的联系，也就便于直接进行地质解释和推断。所以，因子分析已在地质研究工作中得到了广泛的应用。

但是，因子分析还有其不足之处：它把研究变量的R型因子分析与研究样品的Q型因子分析看成两种对立的概念，人为地将两种因子分析割裂开来，这将会漏掉许多有用的信息。事实上，对于同一个地质问题，常常需要同时研究地质成因以及不同类型样品的地质特征，前一个问题要通过对样品的研究，而后一个问题则要通过对变量的研究，才能得到合理的地质解释。这说明R型因子分析与Q型因子分析，是同一问题的不可分割的两个部分。

另外，一般来说，样品的数目远远大于变量的数目。假设有200个样品，每个样品有10个变量，那么，对于R型因子分析来说，只要计算一个（10×10）阶相关矩阵的特征值和特征向量，而对于Q型因子分析来说，就要计算一个（200×200）阶相似系数矩阵的特征值和特征向量，这对于一般的微型计算机来说，无论是内存还是速度都是难以胜任的。而实际研究工作中所要处理的地质问题，样品数目经常超过200个，这对于进行Q型因子分析是十分困难的。

还有一个问题，就是在一般情况下，只能对变量进行标准化，而对样品则不宜进行标准化处理。可见，标准化处理对于变量和样品是非对等的，这就给寻找R型与Q型因子分析之间的联系带来了困难。

鉴于上述情况，在因子分析的基础上产生了对应分析方法。对应分析的主要好处是可由R型因子分析的结果，很容易地导出Q型因子分析的结果，从而克服了由于样品容量过大，给Q型因子分析在计算上带来的困难。对应分析还可以把R型与Q型因子分析统一起来，把变量和样品同时反映在同一因子平面图形上，便于对变量与样品统一进行地质解释和推断。例如，在图形上邻近的一些变量点，显然密切相关，可能指示某一特定的地质作用或地质过程；类似地，在图形上邻近的一些样品点也必然密切相关，可能是同一地质过程的产物，也可能同属一种成因类型。对于同属一种类型的样品点，通常可由邻近它们的变量所表征，这就有助于对样品点的类型进行地质解释，同时由样品的空间分布，又可以了解地质过程的空间关系。

第一节　原始数据的变换

如果有n个样品，每个样品有m个变量，其原始资料矩阵为

$$X=[x_{ij}]_{m\times n}=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix}$$

这里要求$x_{ij}>0$，即x_{ij}应为有实际意义的非负数据。

首先要对原始数据矩阵X按行、按列分别求和以及求全部数据的总和：

$$\begin{array}{cccc|c} x_{11} & x_{12} & \cdots & x_{1n} & \sum\limits_{j=1}^{n} x_{1j}=x_{1.} \\ x_{21} & x_{22} & \cdots & x_{2n} & \sum\limits_{j=1}^{n} x_{2j}=x_{2.} \\ \cdots & \cdots & \cdots & \cdots & \cdots \quad \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} & \sum\limits_{j=1}^{n} x_{mj}=x_{m.} \\ \hline \sum\limits_{i=1}^{m} x_{i1} & \sum\limits_{i=1}^{m} x_{i2} & \cdots & \sum\limits_{i=1}^{m} x_{in} & \sum\limits_{i=1}^{m}\sum\limits_{j=1}^{n} x_{ij}=T \\ \| & \| & & \| & \\ x_{.1} & x_{.2} & & x_{.n} & \end{array}$$

其中

$$x_{.j}=\sum_{i=1}^{m} x_{ij} \quad (j=1,2,\cdots,m)$$

$$x_{i.}=\sum_{j=1}^{n} x_{ij} \quad (i=1,2,\cdots,n)$$

$$T=\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$$

再用数据总和T去除原始数据矩阵中的每一个元素，使原始数据矩阵X变换为Y：

$$Y=\frac{X}{T}=[y_{ij}]_{m\times n}=\begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix}$$

即$y_{ij}=x_{ij}/T \quad (i=1,2,\cdots,m;\ j=1,2,\cdots,n)$

对于R型因子分析来说，是在m维空间中用坐标向量

$$\left[\frac{y_{1j}}{y_{.j}},\ \frac{y_{2j}}{y_{.j}},\ \cdots,\ \frac{y_{mj}}{y_{.j}}\right] \quad (j=1,\ 2,\ \cdots,\ n)$$

的n个点表示n个样品，而每一个样品点的坐标是在该样品中各个变量的相对比例。经此变换后，对于n个样品的研究则可变为对n个样品点相对关系的研究。为了表示任意两个样品点的相关关系，可以引入欧氏距离

$$D_{ab}=\sqrt{\sum_{i=1}^{m}\left(\frac{y_{ia}}{y_{\cdot a}}-\frac{y_{ib}}{y_{\cdot b}}\right)^2}.$$

考虑到各个指标在全部样品中所占的比例不一，在计算样品间的距离时显然会抬高那些数值较大的变量的作用，为消除这种干扰可采用加权距离公式

$$D_{ab}^{*}=\sqrt{\sum_{i=1}^{m}\left(\frac{y_{ia}}{y_{\cdot a}}-\frac{y_{ib}}{y_{\cdot b}}\right)^2\frac{1}{y_{i\cdot}}}$$

$$=\sqrt{\sum_{i=1}^{m}\left(\frac{y_{ia}}{y_{\cdot a}\sqrt{y_{i\cdot}}}-\frac{y_{ib}}{y_{\cdot b}\sqrt{y_{i\cdot}}}\right)^2}.$$

如果继续使用欧氏距离，则每个样品可看作具有如下坐标的向量

$$y_j=\left[\frac{y_{1j}}{y_{\cdot j}\sqrt{y_{1\cdot}}},\ \frac{y_{2j}}{y_{\cdot j}\sqrt{y_{2\cdot}}},\ \cdots,\ \frac{y_{mj}}{y_{\cdot j}\sqrt{y_{m\cdot}}}\right]\quad(j=1,\ 2,\ \cdots,\ n).$$

为了进行因子分析，这里需要求出样品点的方差、协方差矩阵，因而需要求出样品点中第i个变量的均值。

$$\sum_{j=1}^{n}\frac{y_{ij}}{y_{\cdot j}\sqrt{y_{i\cdot}}}y_{\cdot j}=\frac{1}{\sqrt{y_{i\cdot}}}\sum_{j=1}^{n}y_{ij}=\frac{y_{i\cdot}}{\sqrt{y_{i\cdot}}}=\sqrt{y_{i\cdot}}\qquad(i=1,\ 2,\ \cdots,\ m).$$

这一平均值不是算术平均值，而是按概率$y_{\cdot j}$计算的平均值。

对于第i个变量与第j个变量的协方差应为

$$\sum_{k=1}^{n}\left(\frac{y_{ik}}{y_{\cdot k}\sqrt{y_{i\cdot}}}-\sqrt{y_{i\cdot}}\right)\left(\frac{y_{jk}}{y_{\cdot k}\sqrt{y_{j\cdot}}}-\sqrt{y_{j\cdot}}\right)y_{\cdot k}$$

$$=\sum_{k=1}^{n}\left(\frac{y_{ik}}{\sqrt{y_{i\cdot}}\sqrt{y_{\cdot k}}}-\sqrt{y_{i\cdot}}\sqrt{y_{\cdot k}}\right)\left(\frac{y_{jk}}{\sqrt{y_{j\cdot}}\sqrt{y_{\cdot k}}}-\sqrt{y_{j\cdot}}\sqrt{y_{\cdot k}}\right)$$

$$=\sum_{k=1}^{n}\left(\frac{y_{ik}-y_{i\cdot}y_{\cdot k}}{\sqrt{y_{i\cdot}y_{\cdot k}}}\right)\left(\frac{y_{jk}-y_{j\cdot}y_{\cdot k}}{\sqrt{y_{j\cdot}y_{\cdot k}}}\right)=\sum_{k=1}^{n}w_{ik}w_{jk}$$

$$=a_{ij}$$

这里记

$$A=[a_{ij}]_{m\times m}=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1m}\\ a_{21} & a_{22} & \cdots & a_{2m}\\ \cdots & \cdots & \cdots & \cdots\\ a_{m1} & a_{m2} & \cdots & a_{mm}\end{bmatrix}$$

$$W=[w_{ij}]_{m\times n}=\begin{pmatrix}w_{11} & w_{12} & \cdots & w_{1n}\\ w_{21} & w_{22} & \cdots & w_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ w_{m1} & w_{m2} & \cdots & w_{mn}\end{pmatrix}$$

那么$A=WW'$

由于 $w_{ik}=\frac{y_{ik}-y_{i.}y_{.k}}{\sqrt{y_{i.}y_{.k}}}=\frac{\frac{x_{ik}}{T}-\frac{x_{i.}}{T}\frac{x_{.k}}{T}}{\sqrt{\frac{x_{i.}}{T}\frac{x_{.k}}{T}}}=\frac{x_{ik}-x_{i.}x_{.k}/T}{\sqrt{x_{i.}x_{.k}}}$

如果令

$$z_{ik}=\frac{x_{ik}-x_{i.}x_{.k}/T}{\sqrt{x_{i.}x_{.k}}} \qquad (2\text{-}6\text{-}1)$$

则 $W=[w_{ij}]_{m\times n}=Z=[z_{ij}]_{m\times n}$

$$A=ZZ' \qquad (2\text{-}6\text{-}2)$$

经过上述的数据变换，就可以由矩阵A进行R型因子分析。因子轴是A的特征向量与对应的特征值方根的乘积，即

$$F_k=[u_{1k},\ u_{2k},\ \cdots,\ u_{mk}]'\sqrt{\lambda_k} \qquad (k=1,\ 2,\ \cdots,\ p)$$

式中的λ_k为矩阵A的特征值，$[u_{1k},\ u_{2k},\ \cdots,\ u_{mk}]'$为$\lambda_k$的特征向量，而$\lambda_k$为第$k$个因子在总方差中的贡献，即在总方差中所占的比例。

Q型因子分析与R型相类似，可在n维空间中用m个坐标向量

$$y_i=\left[\frac{y_{i1}}{y_{i.}\sqrt{y_{.1}}},\ \frac{y_{i2}}{y_{i.}\sqrt{y_{.2}}},\ \cdots,\ \frac{y_{im}}{y_{i.}\sqrt{y_{.m}}}\right]$$

表示m个变量点，并且可以求出任意两样a，b之间的协方差矩阵

$$B=[b_{ij}]_{n\times n}=\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}$$

其中

$$\begin{aligned} b_{ij}&=\sum_{k=1}^{m}\left(\frac{y_{ki}}{y_{k.}\sqrt{y_{.i}}}-\sqrt{y_{.i}}\right)\left(\frac{y_{kj}}{y_{k.}\sqrt{y_{.j}}}-\sqrt{y_{.j}}\right)y_{k.} \\ &=\sum_{k=1}^{m}\left(\frac{y_{ki}}{\sqrt{y_{.i}}\sqrt{y_{k.}}}-\sqrt{y_{.i}}\sqrt{y_{k.}}\right)\left(\frac{y_{kj}}{\sqrt{y_{.j}}\sqrt{y_{k.}}}-\sqrt{y_{.j}}\sqrt{y_{k.}}\right) \\ &=\sum_{k=1}^{m}\left(\frac{y_{ki}-y_{k.}y_{.i}}{\sqrt{y_{.i}y_{k.}}}\right)\left(\frac{y_{kj}-y_{k.}y_{.j}}{\sqrt{y_{.j}y_{k.}}}\right) \\ &=\sum_{k=1}^{m}z_{ki}z_{kj} \end{aligned}$$

从而有 $B=Z'Z$ $\qquad (2\text{-}6\text{-}3)$

由上述结果可见，A与B之间存在着对应关系，而且从原始数据x_{ij}变换成z_{ij}之后，z_{ij}对于i，j是对等的，即z_{ij}对变量和样品具有对等性。

可以证明，A与B的非零特征根相同。因而，可从R型因子分析出发直接获得Q型因子分析的结果，这样也就克服了由于样品数量过大所带来的Q型因子分析在计算上的困难。此外，由于A与B有相同的特征值，这些特征值表示各个因子所提供的方差，因而在变量空间中的第一因子、第二因子、……，直到第k个因子与样品空间中相应的各个因子在总方差中所占

的百分比完全相同。即从几何学的角度来看，变量空间中各样品点与各因子轴的距离，和样品空间中各变量点与相对应的各因子轴的距离完全相等。因此，可用相同的因子轴同时表示变量和样品，这样就把R型与Q型因子分析统一起来，即可以把变量和样品统一反映在一个因子平面上。

第二节 对应分析的计算步骤

对应分析的原始数据矩阵是由n个样品、每个样品有m个变量的（$m\times n$）个元素组成，即

$$X=[x_{ij}]_{m\times n}$$

其中的x_{ij}表示第j个样品的第i个变量。

首先对X按行、按列分别求和，即

$$x_{i.}=\sum_{j=1}^{n}x_{ij}\quad(i=1,\ 2,\ \cdots,\ m)$$

$$x_{.j}=\sum_{i=1}^{m}x_{ij}\quad(j=1,\ 2,\ \cdots,\ n)$$

再求数据矩阵的总和T

$$\mathbf{T}=\sum_{i=1}^{m}\ \sum_{j=1}^{n}x_{ij}$$

再将X变换为矩阵Z

$$Z=[z_{ij}]_{m\times n}$$

$$z_{ij}=\frac{x_{ij}-x_{i.}x_{.j}/T}{\sqrt{x_{i.}x_{.j}}}$$

一、R型因子分析

1．求乘积矩阵ZZ'的特征值

通常是用雅可比法求得

$$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$$

按其累积百分比

$$\left(\sum_{k=1}^{p}\lambda_i/\sum_{k=1}^{m}\lambda_k\right)\geqslant 85\%$$

取其前p个特征值λ_1，λ_2，…，λ_p，并计算它们对应的单位特征向量u_1，u_2，…，u_p，最后得到因子载荷矩阵

$$F=\begin{pmatrix}u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1p}\sqrt{\lambda_p}\\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2p}\sqrt{\lambda_p}\\ \cdots & \cdots & \cdots\ \cdots & \cdots\\ u_{m1}\sqrt{\lambda_1} & u_{m2}\sqrt{\lambda_2} & \cdots & u_{mp}\sqrt{\lambda_p}\end{pmatrix}$$

2．在因子平面上作变量点图

就是分别在F_1F_2，F_1F_3，…，$F_{(p-1)}F_p$各对因子轴平面上把变量点表示出来。

二、Q型因子分析

1. 求Q型因子载荷矩阵

由R型因子分析所得到的P个特征值$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$,计算其对应于乘积矩阵$Z'Z$的单位特征向量$Z'u_1 = v_1$，$Z'u_2 = v_2$，…，$Z'u_p = v_p$，从而得Q型因子载荷矩阵

$$E=\begin{pmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \cdots & v_{1p}\sqrt{\lambda_p} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \cdots & v_{2p}\sqrt{\lambda_p} \\ \cdots & \cdots & \cdots & \cdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} & \cdots & v_{np}\sqrt{\lambda_p} \end{pmatrix}$$

2. 在因子平面上作样品点图

就是在前面R型因子分析的F_1F_2，F_1F_3，…，$F_{(p-1)}F_p$各个因子轴平面上把样品点表示出来。

三、算　例

[1]东濮凹陷的18个砂岩分析数据见表2-6-1。以砂岩的三种主要碎屑成分即石英、长石及岩块的百分含量为地质变量。共选8口井的18个样品，样品均有人工鉴定的定名结果，通过对应分析检查定名结果是否正确。

表2-6-1　18个砂岩样品的分析数据及岩石定名

样品号	石英含量(%)	长石含量(%)	岩块含量(%)	岩石鉴定定名
1	82	6	9	石英砂岩
2	83	8	9	石英砂岩
3	86	7	5	石英砂岩
4	88	8	12	硬砂质石英砂岩
5	78	7	12	硬砂质石英砂岩
6	84	5	11	硬砂质石英砂岩
7	76	9	13	硬砂质石英砂岩
8	61	26	13	长石砂岩
9	60	25	16	长石砂岩
10	58	29	13	长石砂岩
11	57	27	16	长石砂岩
12	55	30	16	长石砂岩
13	47	27	26	混合砂岩
14	46	25	29	混合砂岩
15	58	7	35	硬砂岩
16	53	17	30	长石质硬砂岩
17	61	12	25	长石质硬砂岩
18	50	17	33	长石质硬砂岩

经过计算，得到3个变量及18个样品的第一、第二主因子载荷值，见表2-6-2；变量及样品的第一、第二主因子载荷平面图，见图2-6-1。

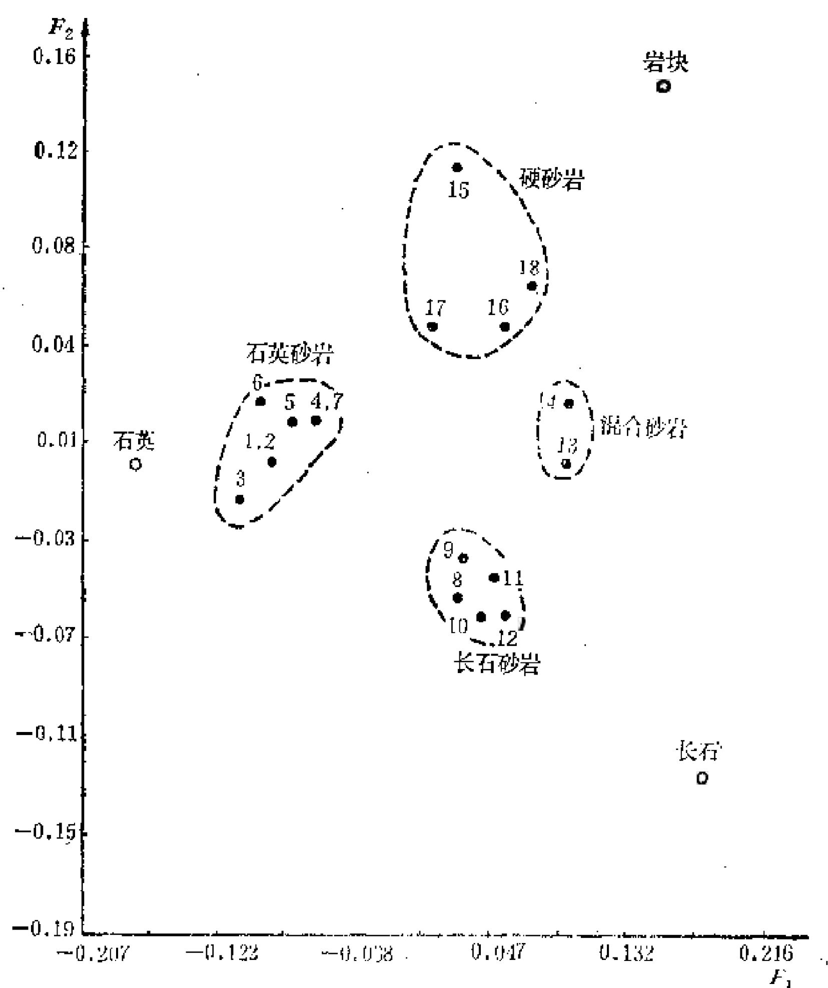


图2-6-1　F_1—F_2主因子载荷平面图

表2-6-2　F_1及F_2因子载荷表

变量或样品	序　　号	F_1	F_2
变　　量	1	−0.1717	−0.0091
	2	0.1811	−0.1313
	3	0.1552	0.1431
样　　品	1	−0.0849	−0.0068
	2	−0.0862	−0.0059
	3	−0.1078	−0.0190
	4	−0.0722	0.0068
	5	−0.0755	0.0112
	6	−0.0928	0.0140
	7	−0.0595	0.0076
	8	0.0278	−0.0673
	9	0.0318	−0.0451

续表

变量或样品	序　号	F_1	F_2
样　品	10	0.0437	−0.0687
	11	0.0470	−0.0484
	12	0.0583	−0.0641
	13	0.0936	−0.0060
	14	0.0969	0.0144
	15	0.0295	0.1083
	16	0.0592	0.0490
	17	0.0130	0.0476
	18	0.0731	0.0618

由图2-6-1可见，三个变量构成的三角形近于等边三角形，18个样品聚集成4个集团，分别属于4种类型的砂岩。而在石英砂岩集团中，还可细分为石英砂岩（样品1，2，3）和硬砂质石英砂岩（样品4，5，6，7）两类；在硬砂岩集团中，也可以区分出硬砂岩（样品15）和长石质硬砂岩（样品16，17，18）；另外两个集团是混合砂岩（样品13，14）以及长石砂岩（样品8，9，10，11，12）。

对应分析的计算结果与人工鉴定结果完全吻合，而因子载荷图实际上相当于划分岩性的三角坐标图。这个结果说明，用对应分析研究成因分类是非常有效的。

表2-6-3　盐泉水化学分析化验数据

样品序号	矿化度 (g/l)	$\frac{Br}{Cl}10^3$	$\frac{K}{\Sigma盐}10^3$	$\frac{K}{Cl}10^3$	$\frac{Na}{K}$	$\frac{Mg}{Cl}10^2$	$\frac{eNa}{eCl}$
1	11.835	0.48	14.36	26.21	25.21	0.81	0.98
2	45.596	0.526	13.85	24.04	26.01	0.91	0.96
3	3.525	0.086	24.4	49.5	11.30	6.82	0.85
4	2.681	0.37	13.57	25.12	26.00	0.82	1.01
5	46.287	0.386	14.6	25.9	23.32	2.18	0.93
6	17.956	0.28	9.75	17.66	37.2	0.464	0.98
7	7.370	0.506	13.6	24.28	10.69	8.8	0.56
8	4.223	0.24	3.8	7.1	88.2	1.11	0.97
9	6.442	0.19	4.7	9.1	75.2	0.74	1.03
10	16.234	0.39	3.1	5.4	121.5	0.42	1.00
11	10.565	0.42	2.4	4.7	135.6	0.87	0.98
12	23.535	0.23	2.6	4.6	141.8	0.31	1.02
13	5.392	0.12	2.8	6.2	111.2	1.14	1.07
14	284.149	0.148	1.763	2.969	215.86	0.140	0.98
15	316.604	0.317	1.453	2.432	263.41	0.249	0.98
16	307.310	0.173	1.627	2.729	235.70	0.214	0.98
17	322.515	0.312	1.382	2.320	282.21	0.024	1.00
18	254.580	0.297	0.599	1.476	410.30	0.239	0.93
19	304.052	0.223	0.789	1.257	438.33	0.193	1.01
20	202.446	0.042	0.741	1.206	305.77	0.29	0.99

〔2〕云南某地钾盐矿泉水化学分析数据见表2-6-3。进行对应分析的目的是为了解样品与变量之间的相互关系，以便更合理地解释各盐泉和变量之间的成因联系。

经过计算，ZZ'矩阵的特征值及累积百分比见表2-6-4。其前两个特征值λ_1，λ_2所代表的方差已占总方差的96.05%，因此，前两个主因子已能够较好地代表整个数据的变化。

表2-6-4　各个因子的特征值累积百分比

因子序号	特征值(λ)	累积值	累积百分比(%)
1	0.4432	0.4432	79.14
2	0.0947	0.5379	96.05
3	0.02096	0.55886	99.79
4	0.00091	0.55977	99.95
5	0.00022	0.55999	99.998
6	0.67×10^{-4}	0.56006	100
7	0.7523×10^{-10}	0.56006	100

取前两个特征值λ_1，λ_2以及相应的前两个特征向量u_1，u_2计算R型因子载荷，得到前两个主因子载荷值，见表2-6-5。

取前两个特征值λ_1，λ_2计算v_1，v_2得到Q型前两个主因子载荷值，见表2-6-6。

由表2-6-5及表2-6-6可以绘制样品及变量的因子平面图，见图2-6-2。

表2-6-5　R型前两个因子载荷表

变量序号	变　量	F_1	F_2
1	矿化度(g/l)	−0.1539	0.2291
2	(Br/Cl)10^3	0.03326	−0.0085
3	(K/Σ盐)10^3	0.3437	0.0129
4	(K/Cl)10^3	0.4989	0.0179
5	Na/K	−0.1087	−0.2025
6	(Mg/Cl)10^2	0.1944	0.0061
7	εNa/εCl	0.0421	−0.0238
方　差　贡　献		0.4432	0.0947
累积方差贡献(%)		79.14	96.05

在图2-6-2中，由于第一因子轴(F_1)的方差贡献已达到79.14%，因而F_1是本区地质因素中占主导地位的一个因子，它反映了该区沉积环境演变的主要特征。在图2-6-2中可以看出F_1的左端为钠，右端为钾，含钾盐泉的主要特征变量(K/Cl)10^3(K/Σ盐)10^3、(Mg/Cl)10^2、εNa/εCl、(Br/Cl)10^3都位于F_1轴的右端，严格受F_1控制。F_1因子载荷中所占比重最大的变量是(K/Cl)10^3(0.4989)，位于F_1轴的最右端，其他特征系数，按其因子载荷在F_1中所占比重的大小自右而左依次排列为：(K/Cl)10^3(0.4989)、(K/Σ盐)10^3(0.3437)、(Mg/Cl)10^2(0.1944)、εNa/εCl(0.0421)、(Br/Cl)10^3(0.03326)、Na/K(−0.1087)。这说明F_1是各种盐类物质随着沉积环境的改变而开始

表2-6-6　Q型前两个因子载荷值

样品序号	盐泉类型	判别分析分类	E_1	E_2
1	钾盐泉	A	0.1993	0.00099
2	钾盐泉	A	0.1446	0.0501
3	钾盐泉	A	0.3841	0.0203
4	钾盐泉	A	0.2103	−0.0157
5	钾盐泉	A	0.1591	−0.0578
6	钾盐泉	A	0.1190	−0.0092
7	钾盐泉	A	0.3070	0.0208
8	钠钾过渡泉	C	0.0297	−0.0946
9	钠钾过渡泉	C	0.0461	−0.0764
10	钠钾过渡泉	C	−0.00023	−0.0982
11	钠钾过渡泉	C	−0.0045	−0.1169
12	钠钾过渡泉	C	−0.0128	−0.1006
13	钠钾过渡泉	C	0.0125	−0.1095
14	钠盐泉	B	−0.0790	0.0844
15	钠盐泉	B	−0.0882	0.0765
16	钠盐泉	B	−0.0849	0.0871
17	钠盐泉	B	−0.0907	0.0695
18	钠盐泉	B	−0.0904	−0.0351
19	钠盐泉	B	−0.0981	−0.0169
20	钠盐泉	B	−0.0792	−0.0231
方差贡献			0.4432	0.0947
累积方差贡献(%)			79.14	96.05

沉积的先后次序，钾的浓度自左向右递增，而钠的浓度递减。因此，可以认为F_1是反映盐类沉积顺序的分异轴。钠的特征系数位于F_1轴的左端，反映了富Na的沉积环境。(Br/Cl) 10^3、(Mg/Cl) 10^2处于F_1轴的中部，说明本区钠盐和钾盐沉积过程中，Mg盐与Br化物具有一定的浓度，而钾盐沉积过程中Mg盐的混入要比Br化物更为明显。

第二因子轴(F_2)的方差贡献比F_1的小的多，仅为16.81%。这一因子轴中，因子载荷所占比重较大的变量是矿化度(0.2291)和Na/K(−0.02025)，分别位于F_2的上、下两端。这两个变量虽然也受F_1的影响，但更主要的是受F_2支配；而其余的变量位于F_2的中部，在F_2上所占因子载荷比重甚小，说明主要是受F_1控制。F_2因子轴主要反映了含盐地层沉积过程中其他矿物质的混入情况，例如Fe_2O_3、硫酸盐类、有机盐的混入等等。这显示了富钠环境的沉积阶段杂质的混入更为明显。

在图2-6-2中，按变量和样品的聚合情况，可以划分为三个区。其中，第Ⅰ区位于F_1轴右端，盐泉特征系数偏于含钾，说明这个区内的样品有利于形成钾盐矿床；第Ⅱ区位于F_1轴中部，该区同时也受F_2的影响，而靠近Na/K系数的一端，盐泉有一定的钾矿化，沉积顺序属于过渡型，盐泉中除有一定浓度的Na以外，其他的混入物也相对增加；第Ⅲ区位于F_1轴的左端，属于钠盐泉，该区样品的F_1因子载荷所占比重都很小，且很相近，都在−0.0792到−0.0981之间，而受F_2的影响最大，F_2因子载荷变化范围很宽，由−0.0231到0.0871(见表2-6-6中的14～20号样品)，这说明钠盐泉受矿化度的影响较大，样品随着矿

化度的变化在F_2轴上呈线状分布。

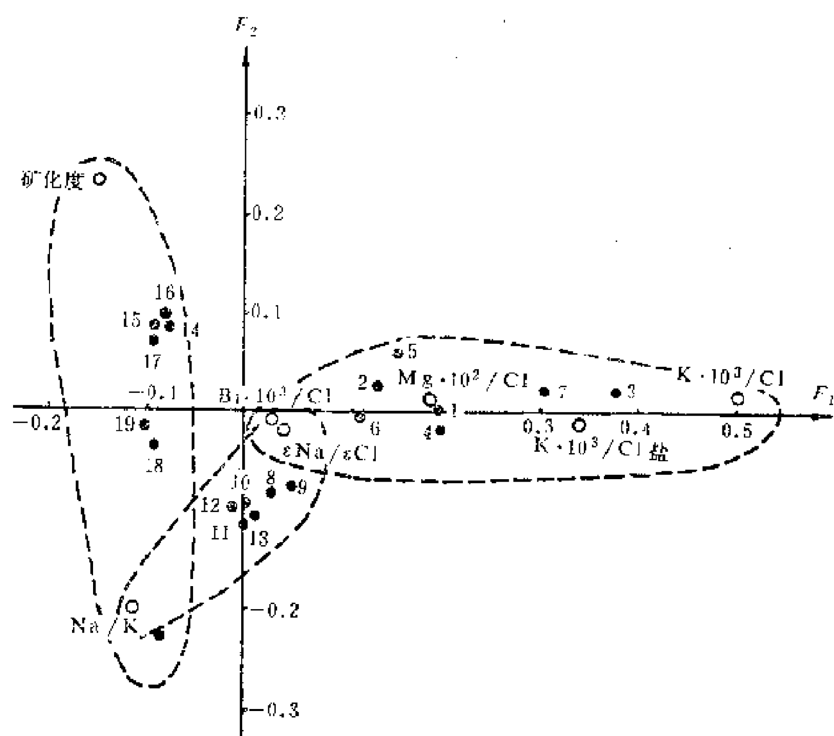


图2-6-2 样品及变量的因子平面图

第七章　非线性映射

非线性映射是一种几何图象降维方法，即通过非线性变换后，把高维空间中的几何图象变换成低维（一维、二维或三维）空间中的图象，并要求变换后仍能近似地保持原象的几何关系。这种方法直观形象，能在低维空间中看到高维空间中样品点之间相互关系的近似图象。

从几何学角度看因子分析也是一种降维方法，R型因子分析是通过线性组合把原有的m个变量点综合成p（$p<m$）个公共因子，以达到简化问题的目的。但是，这种线性变换是有局限性的，例如，在二维平面上有一组样品点呈"S"形分布，见图2-7-1。经过线性组合，两个主分量的方向为F_1和F_2，而样品点无论在F_1方向上还是在F_2方向上都有不可忽略的的方差。可见，经过因子分析后样品点仍为二维分布。但是，如果能够排除干扰，这些样品点也是可以降为一维的。

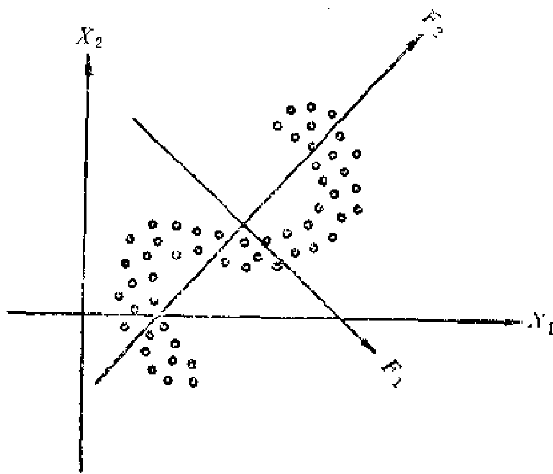


图2-7-1　样品点在平面上的分布

除此之外，许多地质问题若仅用前面两个或三个主要因子进行分析，由于它们所代表的变量方差在变量总方差中所占比例较小，而不能真实地说明问题，也就是说，在这种情况下通过因子分析只能得出一个近似的结果。

除因子分析外，聚类分析也是研究高维空间点群结构的方法。但是，由聚类分析得到的二维谱系图，往往使原有点群之间的结构关系已发生了相当大的畸变。

基于上述原因，很需要用非线性的几何降维方法，去研究高维空间中的点群结构。1969年赛孟（J.W.Sammon）首先提出了非线映射方法，这种方法在一定程度上克服了因子分析和聚类分析方法的不足。

第一节　Q型非线性映射

假定有n个样品，每个样品有m个变量，每个样品可看作为m维空间中的一个点，即

$$X_i=[x_{i1},\ x_{i2},\ \cdots,\ x_{im}]\quad (i=1,\ 2,\ \cdots,\ n).$$

空间中的第i个点与第j个点的欧氏距离为

$$d_{ij}^*=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}\qquad (i,\ j=1,\ 2,\ \cdots,\ n)$$

赛孟提出的非线性映射方法是把n个样品点变换到一个维数较低的P维空间中，变换时尽

可能保持n个点之间的欧氏距离"不变"，特别是要尽可能保持距离比较小的那些点之间的距离不变，而对原来点与点之间距离较大者允许有相对的歪曲。由于受到维数P的限制，因而变换后不可能使所有点之间的距离完全不变，因而，非线性映射只能局部地保持样品点之间的距离不变。

要将m维空间中的n个点，映射到p（$p<m$）维空间中，为了便于作图表示点群结构，一般取低维空间的维数$p=1$，2，或3。通过非线性映射后的样品点为

$$Y_i=[y_{i1},\ y_{i2},\ \cdots,\ y_{ip}]\quad (i=1,\ 2,\ \cdots,\ n)$$

然后计算这n个点之间的欧氏距离$d_{ij}^{(0)}$。

为了达到由高维空间降维到低维空间时，样品点之间距离尽量保持"不变"的目的。需要引入变换的约束条件，或者称误差函数，即

$$E^{(0)}=\frac{1}{\sum\limits_{i<j}^{n}[d_{ij}^{*}]}\sum_{i<j}^{n}\frac{(d_{ij}^{*}-d_{ij}^{(0)})^2}{d_{ij}^{*}} \tag{2-7-1}$$

误差函数E表示低维空间中n个点的构形与高维空间中n个点的构形之间的拟合程度。（2-7-1）式中

$$\sum_{i<j}^{n}[d_{ij}^{*}]=\sum_{i=1}^{n-1}\ \sum_{j=i+1}^{n}[d_{ij}^{*}]$$

是矩阵

$$D=\begin{pmatrix} d_{11}^{*} & d_{12}^{*} & \cdots & d_{1n}^{*} \\ d_{21}^{*} & d_{22}^{*} & \cdots & d_{2n}^{*} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1}^{*} & d_{n2}^{*} & \cdots & d_{nn}^{*} \end{pmatrix}$$

中的上三角（不包括对角线）部分的元素之和。（2-7-1）式中的

$$\frac{(d_{ij}^{*}-d_{ij}^{(0)})^2}{d_{ij}^{*}}$$

即使在E较小时，如果距离d_{ij}^{*}较大，则其与相应的变换后距离$d_{ij}^{(0)}$仍会有明显的差异。因而，同时引入比例因子

$$\frac{1}{\sum\limits_{i<j}^{n}[d_{ij}^{*}]}$$

其目的就是为了比较不同情况下的映射误差。

非线性映射的计算过程是调整n个点$Y_i^{(0)}$（$i=1,\ 2,\ \cdots,\ n$）的坐标$y_{mq}^{(0)}$（$m=1,\ 2,\ \cdots,\ n;\ q=1,\ 2,\ \cdots,\ m$），使其误差尽量减少。假如经过调整后，$n$个点的映象为$Y_i^{(1)}$，则用调正后的坐标$y_{mq}^{(1)}$代替原来的坐标$y_{mq}^{(0)}$，此时就可以算出调整后的构形误差$E^{(1)}$。这种调整过程可以逐步进行，直到误差函数$E$的值达到极小为止。可见，非线性映射的计算是个迭代过程，这里采用的是最速下降方法。其具体计算步骤如下：

如果经过第t步调整后坐标为$y_{mq}^{(t)}$（$m=1,\ 2,\ \cdots,\ n;\ q=1,\ 2,\ \cdots,\ p$），同时得到误差函数$E^{(t)}$。在第（$t+1$）步调整中，映象$Y_i^{(t+1)}$的坐标计算如下：

$$y_{mq}^{(t+1)}=y_{mq}^{(t)}-MF\Delta_{mq}^{(t)} \tag{2-7-2}$$

上式中的MF称为魔力因子，用于控制收敛速度。魔力因子的经验取值为0.3～0.4。$\Delta_{mq}^{(t)}$等于下面的一阶偏导数与二阶偏导数之商，即

$$\Delta_{mq}^{(t)}=\frac{\dfrac{\partial E^{(t)}}{\partial y_{mq}^{(t)}}}{\dfrac{\partial^2 E^{(t)}}{\partial y_{mq}^{2(t)}}} \tag{2-7-3}$$

上式中

$$\frac{\partial E}{\partial y_{mq}}=-\frac{2}{\sum\limits_{i<j}^{n}[d_{ij}^*]}\sum_{\substack{j=1\\ j\neq m}}^{n}\left[\frac{d_{mj}^*-d_{mq}}{d_{mj}^*d_{mj}}\right](y_{mq}-y_{jq}) \tag{2-7-4}$$

$$\frac{\partial^2 E}{\partial y_{mq}^{2}}=-\frac{2}{\sum\limits_{i<j}^{n}[d_{ij}^*]}\sum_{\substack{j=1\\ j\neq m}}^{n}\frac{1}{d_{mj}^*d_{mj}}\left[(d_{mj}^*-d_{mj})-\frac{(y_{mq}-y_{jq})^2}{d_{mj}}\left(1+\frac{d_{mj}^*-d_{mj}}{d_{mj}}\right)\right] \tag{2-7-5}$$

非线性映射的迭代计算过程要进行到使E小于事先指定的精度为止，或者迭代次数达到预先确定的次数为止。

此外，还有两个问题需要说明，第一个问题是非线性映射的计算结果与迭代计算开始时初始点的选择有关。如果初始点选择不当，有可能使误差函数陷入局部最小值，所以，在实际应用时，经常先进行因子分析，选择前面的p个主分量的得分作为初始坐标，因为它们包含了原始n个点的主要信息。第二个问题是非线性映射在原则上可以选择任何定义的距离去度量，但是，对于不同的度量标准，迭代公式也需要作相应的修改。

第二节 R型非线性映射

非线性映射也有Q型与R型之分。基于赛孟提出的Q型非线性映射的思路，有人提出研究变量的R型非线性映射方法。

如果在n维空间有m个变量点，即

$$X_i=[x_{1i},\ x_{2i},\ x_{ni}]\quad (i=1,\ 2,\ \cdots,\ m)$$

假定这m个变量都已经过标准化处理，即均值为0，方差为1，那么，变量之间的相关系数可以表示为

$$r_{ij}^*=\frac{1}{n-1}\sum_{k=1}^{n}x_{ki}x_{kj}\quad (i,\ j=1,\ 2,\ \cdots,\ m) \tag{2-7-6}$$

如果这m个变量在h（$h<n$）维空间中的映象是

$$Y_i=[y_{1i},\ y_{2i},\ \cdots,\ y_{hi}]\quad (i=1,\ 2,\ \cdots,\ h)$$

此时变量之间的相关系数可以表示为

$$r_{ij}=\frac{1}{h-1}\sum_{k=1}^{h}y_{ki}y_{kj}$$

并且选择误差函数

$$E=\frac{2}{m(m-1)}\sum_{i<j}^{m}(r_{ij}^{*}-r_{ij})^2 \qquad (2\text{-}7\text{-}7)$$

作为非线性映射的约束条件。E可以解释为变换前的相关矩阵$R^*=[r_{ij}^*]_{m\times m}$与变换后的相关矩阵$R=[r_{ij}]_{m\times m}$的上三角部分（不包括对角线元素）元素之间的平均平方误差。所采用的最速下降迭代公式为

$$y_{qs}^{(t+1)}=y_{qs}^{(t)}-MF\Delta_{qs}^{(t)} \quad (s=1,\ 2,\ \cdots,\ m;\ q=1,\ 2,\ \cdots,\ h) \qquad (2\text{-}7\text{-}8)$$

上式中，$y_{qs}^{(t)}$、$y_{qs}^{(t+1)}$分别表示第t次、第$t+1$次迭代计算结果；MF为魔力因子，用以控制收敛速度，一般可取$MF=0.3\sim0.4$；

$$\Delta_{qs}=\frac{\dfrac{\partial E}{\partial y_{qs}}}{\dfrac{\partial^2 E}{\partial y_{qs}^2}} \qquad (2\text{-}7\text{-}9)$$

$$\frac{\partial E}{\partial y_{qs}}=\frac{2}{h-1}\sum_{\substack{i=1\\ i\neq s}}^{m}y_{qi}(r_{si}-r_{si}^{*}) \qquad (2\text{-}7\text{-}10)$$

$$\frac{\partial^2 E}{\partial y_{qs}^2}=\frac{2}{(h-1)^2}\sum_{\substack{i=1\\ i\neq s}}^{m}y_{qi}^2 \qquad (2\text{-}7\text{-}11)$$

（2-7-10）式中的r_{si}（s，$i=1$，2，…，m）表示根据前一次迭代结果y_{qs}计算而得到的相关系数。

为了选准迭代的初始值$y_{qs}^{(0)}$，可以事先进行R型因子分析并选用前h个主分量的载荷作为$y_{qs}^{(0)}$。如果取$h=2$，则映象可在一张平面图上表示出来，这对研究点群的空间结构是很方便的。

第三节　算　　例

〔1〕原始数据为六维空间的正三角形，由9个点组成，这9个点均匀分布在三条边上，现在要把这些点经过降维映射到二维平面上。

取二维平面上的正方形为初始构形，二维平面上8个初始点均匀分布在4个边上，1个点在正方形的中心。经过迭代计算，当取$E=1.0\times10^{-22}$时，Y_i（$i=1$，2，…，9）的二维映象为图2-7-2。变换前的六维正三角形的坐标值及变换后的二维映象的坐标值见表2-7-1。

由图2-7-2看出，六维正三角形的二维映象仍然是一个正三角形，说明通过非线映射可以直观地在二维空间中看到高维空间的图象。

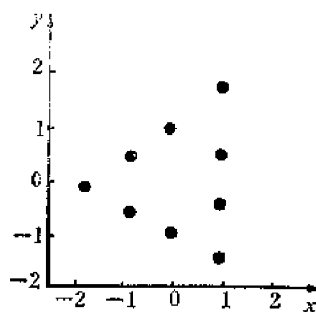


图2-7-2　六维正三角形的二维映象

〔2〕原始数据为六维空间中的球，由9个点组成，其中8个点均匀分布在球面上，一个点在球心上，把它们映象到二维平面上，初始二维构形与前算例相同。经过迭代计算，当取$E=1.7\times10^{-22}$时，Y_i（$i=1$，2，…，9）的二维映象见图2-7-3。变换前的六维球的坐标值及变换后的二维映象的坐标值见表2-7-2。

表2-7-1　六维正三角形的非线性二维映象

样品号＼坐标	六维样品点 1	2	3	4	5	6	二维映象点 1	2
1	0	0	0	0	0	0	0.770	1.542
2	0.370	0.640	−0.429	−0.061	0.466	0.140	−0.067	0.996
3	0.739	1.280	−0.857	−0.122	−0.933	0.280	−0.905	0.450
4	1.109	1.920	−1.286	−0.184	−1.489	0.420	−1.743	−0.095
5	1.172	1.811	−1.225	−0.612	0.687	0.30	−0.852	−0.548
6	1.236	1.701	−1.164	−1.041	−0.116	−0.360	0.040	−1.001
7	1.299	1.591	−1.102	−1.470	0.919	−0.750	0.932	−1.454
8	0.866	1.061	−0.735	−0.980	0.612	−0.500	0.878	0.455
9	0.433	0.530	−0.367	−0.490	0.306	−0.250	0.824	0.543

表2-7-2　六维球的非线性二维映射

样品号＼坐标	六维样品点 1	2	3	4	5	6	二维映象点 1	2
1	0	0	0	0	0	0	0.707	0.707
2	0.252	0.462	−0.308	−0.007	−0.441	0.143	0	1.000
3	0.610	0.963	−0.650	−0.278	−0.444	0.056	−0.707	0.707
4	0.864	1.212	−0.827	−0.686	−0.008	−0.210	−1.000	0
5	0.866	1.061	−0.735	−0.980	0.612	−0.500	−0.707	−0.707
6	0.614	0.599	−0.427	−0.986	1.053	−0.643	0	−1.000
7	0.256	0.097	−0.085	−0.702	1.056	−0.558	0.707	−0.707
8	0.002	−0.151	0.082	−0.293	0.620	−0.290	1.000	0
9	0.433	0.530	−0.367	−0.490	0.306	−0.250	0	0

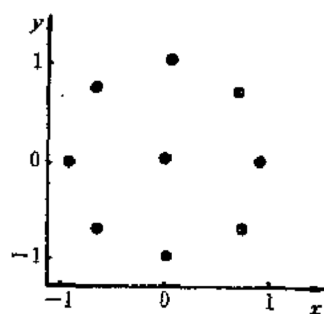


图2-7-3　六维球的二维映象

图2-7-3为六维球面上8个均匀分布点及球心的二维映象图，映象在二维平面上仍然为一个圆，一个点在圆心上，其他8个点均匀分布在圆周上，直观地把高维空间中点的构形映象为低维空间中点的构形。

[3]引用因子分析中的算例资料，即用周口盆地的白垩系、下第三系生油岩抽提物和油样的甾烷有关的地球化学分析数据，进行非线性映射计算。原始数据见表2-5-1，以表2-5-3中的因子得分F_1，F_2作为二维空间的初始值。经过计算得到降维后的映象数值见表2-7-3。33个样品点的二维空间结构见图2-7-4。

在样品点的二维空间结构图上，33个样品点聚为两大集团。第一集团中的大部分样品为

表2-7-3 降维后的平面变量值

样品序号	变量1	变量2	样品序号	变量1	变量2
1	−28.782	−16.093	18	−25.503	−18.142
2	−26.762	−17.840	19	−22.002	−12.161
3	−26.399	−16.536	20	−30.865	−13.909
4	−28.196	−17.804	21	−25.467	−20.098
5	−29.679	−24.066	22	−24.918	−20.774
6	−37.876	−19.406	23	−29.135	−11.386
7	−28.119	−13.270	24	−16.338	−25.946
8	−32.159	−22.712	25	−12.098	−29.801
9	−34.297	−11.399	26	−21.677	−25.728
10	−33.587	−20.476	27	−25.965	−18.063
11	−27.935	−15.456	28	−10.040	−26.948
12	−30.011	− 8.805	29	−19.258	−20.461
13	−28.065	−19.234	30	−30.735	−10.942
14	−24.505	−17.505	31	−21.456	−24.192
15	−20.943	−17.699	32	−13.841	−27.167
16	−28.867	−16.738	33	−13.683	−23.638
17	−29.729	−9.817			

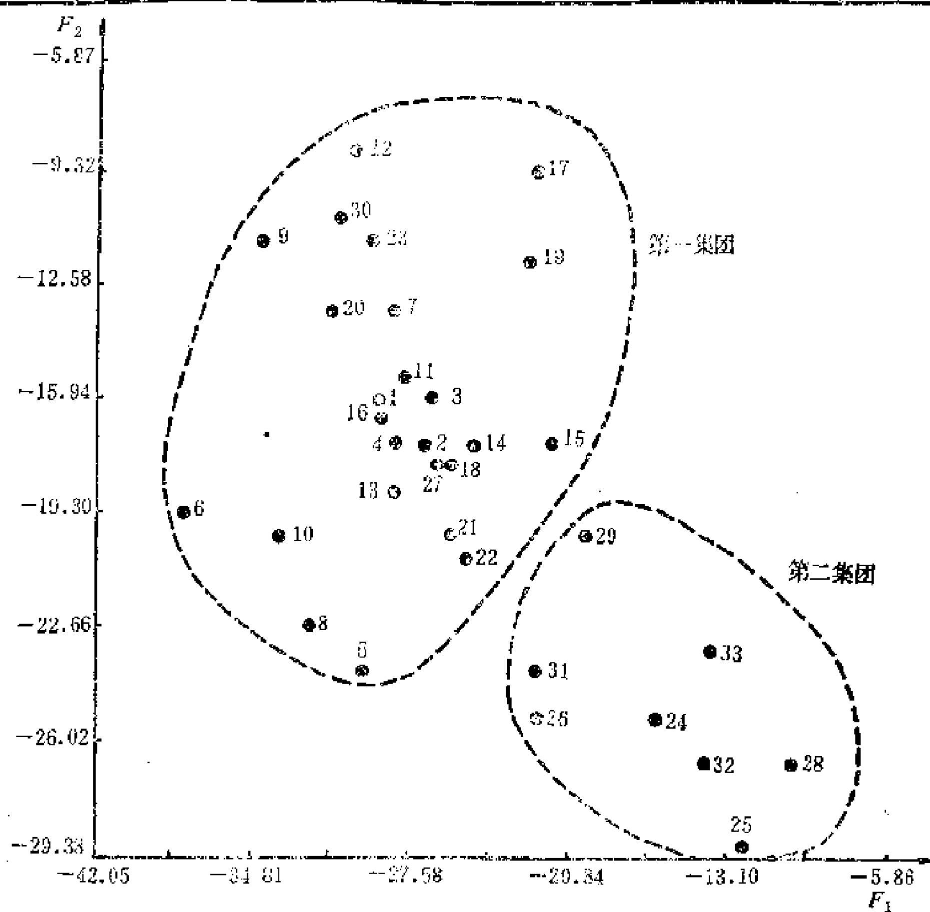


图2-7-4 样品点的二维空间结构图

白垩系生油岩样品，属于母质较差，成熟度较低，经过一定距离运移的类型，少数样品为下第三系样品。据对这些样品的综合研究表明，这些样品已受白垩系生成油气运移的侵染。

第二集团中的样品均为下第三系的生油岩样品，属于母质类型较好，未成熟，且未经运移。

第八章　马尔科夫概型分析

任何一个地质过程都是十分复杂的，既受某些确定性的地质因素控制，也受某些随机性的地质因素支配。也就是说，复杂的地质过程都应当是确定型与随机型两种过程在空间上、时间上不同层次的复合过程。但是，目前在研究地质过程时，往往只注重于确定型过程，而忽视了随机型过程。研究随机型地质过程是目前的一个薄弱环节，到目前为止，所涉及的数学方法还仅限于马尔科夫概型分析，而且其应用领域不广，除个别问题外，多数问题的实际效果也不理想。

1949年，维斯捷列乌斯在研究复式沉积层形成问题时，首先应用了马尔科夫链。本世纪60年代，有关马尔科夫过程在地质学中应用的论文大量出现。70年代以后虽然论文数量有所下降，但人们对其理解程度却在加深，应用领域也有所扩大。

1984年，在莫斯科举行的第27届国际地质会议上，维斯捷列乌斯、阿格特伯格（F.P. Agterberg）等数学地质学家发表了莫斯科宣言，其主要思想是：随机型模型应作为数学地质模型的基础，并且肯定了马尔科夫过程（特别是马尔科夫链）在地质学中应占有特殊的地位。

目前，马尔科夫概型分析主要应用于地层、沉积方面的地质研究，其研究内容有如下三个方面：

（1）用马尔科夫链模拟地层剖面；

（2）证明沉积旋回符合马尔科夫过程；

（3）从沉积机制方面研究地层剖面的演化。

值得一提的是，马尔科夫链在地质学上应用的典型事例是苏联维斯捷列乌斯研究小组自1966年开始的用多重马尔科夫链研究理想花岗岩的形成和演化问题，他们用了整整6年时间才建立起岩浆结晶的精确模型，其后经过多次改进到1984年又提出更为完善的模型。

第一节　马尔科夫过程的含义

如果给定体系的未来状态独立于过去的历史，亦即过程的将来发展完全取决于其现在的状态，而并不依赖于现在状态如何由过去发展而来，也就是说与现在以前的状态无关，体系的这种性质称为“马尔科夫性质”。

如果给定过程的“现在”，其过程的“将来”独立于“过去”，这样的随机过程就称为“马尔科夫过程”。准确地说，该过程的特征为，若$t_1<t_2<\cdots<t_n$为时间参数，且$1<j<n$，则对于给定的$x(t_j)$。随机变量的集合$\{x(t_1), x(t_2),\cdots, x(t_{j-1})\}$和$\{x(t_{j+1}), x(t_{j+2}),\cdots,x(t_n)\}$相互独立；或者等价地说，对于给定的$x(t_1),x(t_2),\cdots, x(t_{n-1})$序列，$x(t_n)$的条件概率分布只依赖于$x(t_{n-1})$的特定值，亦即其概型是在给定$x(t_{n-1})$的条件下，$x(t_n)$的条件概率分布

$$p\{x(t_n)<x_n|x(t_1)=x_1, \cdots, x(t_{n-1})=x_{n-1}\}$$
$$=p\{x(t_n)\leqslant x_n|x(t_{n-1})=x_{n-1}\}$$

可见，一个马尔科夫过程对于最近瞬间之前是无"记忆"的，这个性质称为"无后效性"；因而，马尔科夫过程又称"无后效随机过程"。

按过程的状态和时间性质，马尔科夫过程可分类如下：

（1）状态（随机变量$x(t)$的取值）间断，时间离散的称离散参数马尔科夫链；

（2）状态间断，时间连续的称为连续参数马尔科夫链；

（3）状态连续，时间离散的称为离散参数马尔科夫过程；

（4）状态连续，时间连续的称为连续参数马尔科夫过程。

状态间断的马尔科夫过程总称为"马尔科夫链"。马尔科夫链中，体系的可能状态数目应当是有穷的或可数无穷的。

在研究地质现象时，有时能直接确定其时间上的先后顺序，而有时只能间接地用空间上的上下、前后、左右关系来代替；有时可以找到确定的时间序列，有时只能间接地用距离来代替时间。但是，只要空间序列有类似于马尔科夫性质的关系存在，则仍然可以应用马尔科夫链进行研究。这里将既适用于时间序列又适用于空间序列的马尔科夫概率模型统称为"马尔科夫概型"。

若马尔科夫过程的转移概率随时间而变，则称为非齐次或非平稳马尔科夫过程。而转移概率不随时间而变的马尔科夫过程称为齐次或平稳马尔科夫过程，这是马尔科夫过程的一种特殊类型。目前在地质研究过程中，主要是应用平稳马尔科夫过程。

在地质研究中，如果把岩性看成是一个随机运动着的量，那么，地层剖面中不同的岩性，如砂岩、泥岩、页岩、石灰岩等等可以当成m种不同的状态，并可标以E_1，E_2，…，E_m，岩性每经历一个单位时间则作一次随机转移。设它现在是处于状态E_i，那么下次观察时它可能转移到E_j，也可能转移到E_k。由于事先并不能确切地预言到底转移到哪种状态，而只能给出它转移到某个状态的概率，此即转移概率。基于这种情况，应用转移概率可以对未来时刻出现的状态种类进行预报。此外，马尔科夫链还可用于地层模拟、水系、沉积过程、火山喷发序列等方面的研究。

第二节　马尔科夫链的转移概率

一、一阶转移概率

如果岩性这个量在$t=n$时刻处于状态E_i，在此条件下，下一次即$t=n+1$时刻它转移到状态E_j的概率是

$$p\{x_{n+1}=j|x_n=i\}=p_{ij}$$

这个概率称为一阶转移概率。它实际上是在已知"$x_n=i$"条件下，经过一步转移到"$x_{n+1}=j$"的条件概率。为了明确起见，一般把从状态E_i到E_j的一阶转移概率p_{ij}记为$p_{ij}^{(1)}$，把二阶转移概率记为$p_{ij}^{(2)}$，…，把从状态E_i出发经过t步转移到状态E_j的转移概率记为$p_{ij}^{(t)}$，如果这个t阶转移概率只与状态E_i，E_j以及t有关，而与具体哪个时刻无关，这就是前面提到的齐次或平稳的马尔科夫链。

这种仅与最后状态有直接关系的马尔科夫链可称为一重马尔科夫链；如果一个状态的条件概率不仅与前一个状态，而且与前两个状态，甚至多个状态有关时则称为二重或多重马尔科夫链。

对于马尔科夫链来说，转移概率完全描述了它的概率统计特征，因此，如何确定转移概率，是研究马尔科夫链的一个重要问题。转移概率在理论上是条件概率，而实际应用时是以频率$n_{ij}/n_{i.}$代替"条件概率"来估计转移概率的，即

$$\hat{p}_{ij}=\frac{n_{ij}}{n_{i.}}$$

其中$n_{i.}$是状态E_i出现的次数，n_{ij}是由状态E_i一步转移到状态E_j的次数。

例如，若在某个剖面中岩性这个随机变量只能取砂岩（E_1）、泥岩（E_2）两种状态，下面是某个地层剖面的实际观测记录：

$$E_1E_1E_2E_2E_1E_1E_1E_2E_2E_2E_1E_2E_2E_1E_1E_1$$

某次观测为状态E_1条件下，下次观测为E_1的条件概率记为p_{11}，即下标的第一个数字表示起始状态，第二个数字表示终止状态。上面列出的实际剖面中，出现砂岩（E_1）为9次，最后一次出现砂岩的后面已无资料，所以以砂岩为起始状态来统计下次出现什么状态只能统计8次。经过统计得出 $p_{11}=\frac{5}{8}$， $p_{12}=\frac{3}{8}$。而以泥岩为起始状态来统计下次出现什么状态有7次，经过统计得出 $p_{21}=\frac{3}{7}$， $p_{22}=\frac{4}{7}$。当观测次数很多时，频率接近条件概率，因而可以用频率代替条件概率。所以，下面就把频率当作转移概率进行讨论。把上面统计出的几个转移概率写成矩阵则有

$$\boldsymbol{P}=\begin{bmatrix}p_{11} & p_{12}\\ p_{21} & p_{22}\end{bmatrix}=\begin{pmatrix}\frac{5}{8} & \frac{3}{8}\\ \frac{3}{7} & \frac{4}{7}\end{pmatrix}$$

这个矩阵称为马尔科夫链的转移概率矩阵，这是个一阶转移概率矩阵，记为$p^{(1)}$。这一矩阵的元素为非负的，矩阵中各行元素之和为1。

如果状态不仅有两种，而是有m种，即E_1，E_2，…，E_m，那么由状态E_i经过一步转移到状态E_j的一阶转移概率矩阵为

$$\boldsymbol{P}^{(1)}=[p_{ij}^{(1)}]_{m\times m}=\begin{pmatrix}p_{11}^{(1)} & p_{12}^{(1)} & \cdots & p_{1m}^{(1)}\\ p_{21}^{(1)} & p_{22}^{(1)} & \cdots & p_{2m}^{(1)}\\ \cdots & \cdots & \cdots & \cdots\\ p_{m1}^{(1)} & p_{m2}^{(1)} & \cdots & p_{mm}^{(1)}\end{pmatrix}$$

转移概率矩阵有如下两个性质：

（1） $0\leqslant p_{ij}^{(1)}\leqslant 1$；

（2） $\sum_{j=1}^{m} p_{ij}^{(1)}=1 \quad (i=1, 2, \cdots, m)$

二、高阶转移概率

如果马尔科夫链有m种状态E_1，E_2，…，E_m，从状态E_i出发经两步转移到状态E_j的概率（不管第一步是什么状态）称为二阶转移概率，记为$p_{ij}^{(2)}$，用二阶转移概率排成的矩阵为

$$P^{(2)}=[p_{ij}^{(2)}]_{m\times m}=\begin{pmatrix} p_{11}^{(2)} & p_{12}^{(2)} & \cdots & p_{1m}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} & \cdots & p_{2m}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(2)} & p_{m2}^{(2)} & \cdots & p_{mm}^{(2)} \end{pmatrix}$$

这个矩阵称为二阶转移概率矩阵，其中元素$p_{ij}^{(2)}$可以由实际资料统计出来，即

$$p_{ij}^{(2)}=\frac{E_i\text{后的第二步是}E_j\text{的次数}}{E_i\text{出现的次数}}$$

更一般地，由状态E_i经t步转移到状态E_j的概率$p_{ij}^{(t)}$称为t阶转移概率，其转移概率矩阵为

$$P^{(t)}=[p_{ij}^{(t)}]_{m\times m}=\begin{pmatrix} p_{11}^{(t)} & p_{12}^{(t)} & \cdots & p_{1m}^{(t)} \\ p_{21}^{(t)} & p_{22}^{(t)} & \cdots & p_{2m}^{(t)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(t)} & p_{m2}^{(t)} & \cdots & p_{mm}^{(t)} \end{pmatrix}$$

称为t阶转移概率矩阵，其中

$$p_{ij}^{(t)}=\frac{E_i\text{后第}t\text{步是}E_j\text{的次数}}{E_i\text{出现的次数}}$$

且有性质　$0\leqslant p_{ij}^{(t)}\leqslant 1$，　$\sum_{j=1}^{m}p_{ij}^{(t)}=1$。

对于高阶转移概率的计算，事实上可以利用一阶转移概率根据马尔科夫链的无后效性而得到，例如，对于二阶转移概率

$$\begin{aligned}
p_{ij}^{(2)}&=p\{x_2=j|x_0=i\}\\
&=p\{x_1=1,\ x_2=j|x_0=i\}+p\{x_1=2,\ x_2=j|x_0=i\}+\cdots\\
&\quad+p\{x_1=m,\ x_2=j|x_0=i\}\\
&=\frac{p\{x_0=i,x_1=1,x_2=j\}}{p\{x_0=i\}}+\frac{p\{x_0=i,x_1=2,x_2=j\}}{p\{x_0=i\}}+\cdots\\
&\quad+\frac{p\{x_0=i,x_1=m,x_2=j\}}{p\{x_0=i\}}\\
&=\frac{p\{x_0=i,x_1=1\}}{p\{x_0=i\}}p\{x_2=j|x_0=i,\ x_1=1\}\\
&\quad+\frac{p\{x_0=i,x_1=2\}}{p\{x_0=i\}}p\{x_2=j|x_0=i,\ x_1=2\}+\cdots\\
&\quad+\frac{p\{x_0=i,x_1=m\}}{p\{x_0=i\}}p\{x_2=j|x_0=i,\ x_1=m\}
\end{aligned}$$

$$=p_{i1}^{(1)}p_{1j}^{(1)}+p_{i2}^{(1)}p_{2j}^{(1)}+\cdots+p_{im}^{(1)}p_{mj}^{(1)}$$

$$=\sum_{k=1}^{m}p_{ik}^{(1)}p_{kj}^{(1)} \qquad (2\text{-}8\text{-}1)$$

因而有

$$\boldsymbol{P}^{(2)}=\begin{pmatrix} p_{11}^{(2)} & p_{12}^{(2)} & \cdots & p_{1m}^{(2)} \\ p_{21}^{(2)} & p_{22}^{(2)} & \cdots & p_{2m}^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(2)} & p_{m2}^{(2)} & \cdots & p_{mm}^{(2)} \end{pmatrix}$$

$$=\begin{pmatrix} \sum_{k=1}^{m}p_{1k}^{(1)}p_{k1}^{(1)} & \sum_{k=1}^{m}p_{1k}^{(1)}p_{k2}^{(1)} & \cdots & \sum_{k=1}^{m}p_{1k}^{(1)}p_{km}^{(1)} \\ \sum_{k=1}^{m}p_{2k}^{(1)}p_{k1}^{(1)} & \sum_{k=1}^{m}p_{2k}^{(1)}p_{k2}^{(1)} & \cdots & \sum_{k=1}^{m}p_{2k}^{(1)}p_{km}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{k=1}^{m}p_{mk}^{(1)}p_{k1}^{(1)} & \sum_{k=1}^{m}p_{mk}^{(1)}p_{k2}^{(1)} & \cdots & \sum_{k=1}^{m}p_{mk}^{(1)}p_{km}^{(1)} \end{pmatrix}$$

$$=\begin{pmatrix} p_{11}^{(1)} & p_{12}^{(1)} & \cdots & p_{1m}^{(1)} \\ p_{21}^{(1)} & p_{22}^{(1)} & \cdots & p_{2m}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(1)} & p_{m2}^{(1)} & \cdots & p_{mm}^{(1)} \end{pmatrix}\begin{pmatrix} p_{11}^{(1)} & p_{12}^{(1)} & \cdots & p_{1m}^{(1)} \\ p_{21}^{(1)} & p_{22}^{(1)} & \cdots & p_{2m}^{(1)} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m1}^{(1)} & p_{m2}^{(1)} & \cdots & p_{mm}^{(1)} \end{pmatrix}$$

$$=\boldsymbol{P}^{(1)}\cdot\boldsymbol{P}^{(1)}=(\boldsymbol{P}^{(1)})^2$$

从而，对于高阶转移概率矩阵有

$$\boldsymbol{P}^{(t)}=(\boldsymbol{P}^{(1)})^t$$

更一般的情况是，对于任何r，可以导出

$$p_{ij}^{(t)}=\sum_{k=1}^{m}p_{ik}^{(r)}p_{kj}^{(t-r)} \qquad (2\text{-}8\text{-}2)$$

也就是说，从状态E_i出发经过t步到达状态E_j这一过程，可以看作为它是先经过r（$0<r<t$）步转移到某一状态E_k（$k=1, 2, \cdots, m$），再由E_k经过（$t-r$）步转移到达状态E_j。

第三节　遍历定理与极限分布

马尔科夫链的遍历性的直观意义是：不论从哪个初始状态E_i出发，当转移步数t充分大后，它到达状态E_j的概率是一个不随时间变化的常数p_j。也就是说，无论初始状态如何，经过若干步转移以后，系统将处于平衡状态。因而，反过来，当t充分大时，可用p_j作为$p_{ij}^{(t)}$的近似值。这样，便可以解决当t很大时高阶转移概率的计算问题。p_j称为马尔科夫链的极限概率，而遍历性的中心问题是要确定在什么样的条件下，转移概率的极限才是存在的；极限概率是否构成一个概率分布；以及如何计算极限概率p_j。

遍历性定理是指对于有限状态的马尔科夫链，若存在一个正整数s，使得$p_{ij}^{(s)}>0$对任何i，$j=1, 2, \cdots, m$成立，那么极限

$$\lim_{n\to\infty} p_{ij}^{(n)}=p_j \qquad (2\text{-}8\text{-}3)$$

存在，并且与i无关；而（2-8-3）式中的$\{p_1, p_2, \cdots, p_m\}$是方程组

$$p_j=\sum_{i=1}^{m} p_i p_{ij}^{(1)} \quad (j=1, 2, \cdots, m) \qquad (2\text{-}8\text{-}4)$$

在满足条件$p_j>0$，$\sum_{j=1}^{m} p_j=1$时的唯一解。

例如，有一马尔科夫链，其转移状态有两种：E_1，E_2。经计算得出它的一阶转移概率矩阵为

$$\boldsymbol{P}^{(1)}=\begin{bmatrix}0.79 & 0.21\\ 0.59 & 0.41\end{bmatrix}$$

当$s=1$时，对一切i，j，$p_{ij}^{(1)}>0$满足遍历性定理，故有$\lim_{n\to\infty} p_{ij}^{(n)}=p_j>0$。而$p_j$可由方程组

$$\begin{cases} p_j=\sum_{i=1}^{m} p_i p_{ij}^{(1)} & (j=1, 2, \cdots, m)\\ \sum_{j=1}^{m} p_j=1 & (p_j>0)\end{cases}$$

求出。对于本例为

$$\begin{cases} p_1=0.79p_1+0.59p_2\\ p_2=0.21p_1+0.41p_2\\ p_1+p_2=1 & (p_1, p_2>0)\end{cases}$$

最后得到$p_2=0.26$，$p_1=0.74$。所以，其极限概率矩阵为

$$\widetilde{\boldsymbol{P}}=\begin{bmatrix}0.74 & 0.26\\ 0.74 & 0.26\end{bmatrix}$$

如果从公式$p^{(t)}=(p^{(1)})^t$出发，计算其高阶转移概率有

$$p^{(2)}=p^{(1)}\cdot p^{(1)}=\begin{bmatrix}0.75 & 0.25\\ 0.71 & 0.29\end{bmatrix}$$

$$p^{(3)}=p^{(2)}\cdot p^{(1)}=\begin{bmatrix}0.74 & 0.26\\ 0.74 & 0.26\end{bmatrix}$$

$$p^{(4)}=p^{(3)}\cdot p^{(1)}=\begin{bmatrix}0.74 & 0.26\\ 0.74 & 0.26\end{bmatrix}$$

从各阶转移概率可以看出，其前三阶有所不同，随着阶数增加，3、4阶转移概率矩阵相等，等于极限概率矩阵，而且矩阵中每一列内各元素均相等，即经过若干步转移后，终止状态E_j的概率是一个常数p_j。这就是状态E_j的极限概率。

第四节　马尔科夫概型检验

任何一个系统状态$\{x_i, i=0, 1, 2, \cdots\}$都可以构成一个转移概率矩阵。但是，它是否具有马尔科夫概型的性质，则必须进行独立性的检验。通常是用χ^2检验。

皮尔逊已经证明统计量

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{m}\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2\Big/\frac{n_{i.}n_{.j}}{n}$$

在独立假设下，当n很大时服从自由度$(m-1)^2$的χ^2分布。其中，m为系统的状态数；n_{ij}为转移频数（由状态E_i经过一步转移到E_j的次数）。

$$n_{i.}=\sum_{j=1}^{m}n_{ij}$$

$$n_{.j}=\sum_{i=1}^{m}n_{ij}$$

$$n=\sum_{i=1}^{m}n_{i.}=\sum_{j=1}^{m}n_{.j}$$

例如，本章第二节地层剖面中砂岩与泥岩的出现次数的状态转移频数如表2-8-1：

表2-8-1　状态转移频数表

x_{t-1}	x_t		
	n_{ij}		$n_{i.}$
	E_1	E_2	
E_1	7	3	10
E_2	3	2	5
$n_{.j}$	10	5	$n=15$

假设H_0：x_t与x_{t-1}是相互独立的，为检这一假设，计算统计量

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{2}\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2\Big/\frac{n_{i.}n_{.j}}{n}$$

$$=\frac{\left(n_{11}-\frac{n_{1.}n_{.1}}{n}\right)^2}{\frac{n_{1.}n_{.1}}{n}}+\frac{\left(n_{12}-\frac{n_{1.}n_{.2}}{n}\right)^2}{\frac{n_{1.}n_{.2}}{n}}$$

$$+\frac{\left(n_{21}-\frac{n_{2.}n_{.1}}{n}\right)^2}{\frac{n_{2.}n_{.1}}{n}}+\frac{\left(n_{22}-\frac{n_{2.}n_{.2}}{n}\right)^2}{\frac{n_{2.}n_{.2}}{n}}$$

$$=\frac{\left(7-\frac{10\times 10}{15}\right)^2}{\frac{10\times 10}{15}}+\frac{\left(3-\frac{10\times 5}{15}\right)^2}{\frac{10\times 5}{15}}$$

$$+\frac{\left(3-\frac{5\times 10}{15}\right)^2}{\frac{5\times 10}{15}}+\frac{\left(2-\frac{5\times 5}{15}\right)^2}{\frac{5\times 5}{15}}$$

由于统计量χ^2只是在n很大时服从自由度为$(m-1)^2$的χ^2分布，而此例中，$m=2$，即自由度为1，必须修正$\frac{n_{i.}n_{.j}}{n}$值，这里是减去0.5。即

$$\chi^2=\frac{\left(7-\frac{20}{3}-0.5\right)^2}{\frac{20}{3}}+\frac{\left(3-\frac{10}{3}-0.5\right)^2}{\frac{10}{3}}$$

$$+\frac{\left(3-\frac{10}{3}-0.5\right)^2}{\frac{10}{3}}+\frac{\left(2-\frac{5}{3}-0.5\right)^2}{\frac{5}{3}}=0.438$$

而$(m-1)^2=(2-1)^2=1$自由度的$\chi^2_{0.05}=3.84$。由于$\chi^2=0.438<3.841$，故接受H_0假设，即认为t_1时刻系统处于什么状态与t时刻系统所处状态无关，所以该例中给出的系统状态并非马尔科夫链。

第五节　算　　例

本算例是根据马尔科夫一阶转移概率矩阵研究沉积旋回（据成都地质学院）。

地层剖面取自某地区的钻孔资料，岩性为紫红色粉砂质泥岩夹长石砂岩透镜体，地层层数共90层（$n=90$）；有五种岩性，即状态分为5种（$m=5$）：E_1（砂砾岩）、E_2（粉砂岩）、E_3（泥页岩）、E_4（砂、粉砂页岩）、E_5（煤、炭质页岩）。

1. 转移频数矩阵为

$N=[n_{ij}]=$

	E_1	E_2	E_3	E_4	E_5	$n_{i.}$
E_1	0	7	2	6	2	17
E_2	7	0	2	12	1	22
E_3	2	1	0	1	3	7
E_4	1	6	1	0	15	23
E_5	6	8	2	4	0	20
$n_{.j}$	16	22	7	23	21	89

2. 求一阶转移矩阵

由n_{ij}可以求得频率转移概率$\hat{p}_{ij}=\frac{n_{ij}}{n_{i.}}$，用它作为理论概率估计值可以得到一阶概率转

移矩阵如下

$$P^{(1)}=[\hat{p}_{ij}]=\begin{pmatrix} 0 & 0.41 & 0.12 & 0.35 & 0.12 \\ 0.32 & 0 & 0.09 & 0.54 & 0.04 \\ 0.28 & 0.14 & 0 & 0.14 & 0.43 \\ 0.04 & 0.26 & 0.04 & 0 & 0.65 \\ 0.30 & 0.40 & 0.10 & 0.20 & 0 \end{pmatrix}$$

3. 由$P^{(1)}$求其极限概率而得到

$$\overline{P}=\begin{pmatrix} 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \\ 0.1826 & 0.2471 & 0.0786 & 0.2576 & 0.2340 \end{pmatrix}$$

4. 简化旋回模式

取出固定向量[0.1826，0.2471，0.0786，0.2576，0.2340]，由其中最大概率的状态作为旋回的开始，因而取E_4作为开始，直接利用$P^{(1)}$来画出旋回模式，见图2-8-1。

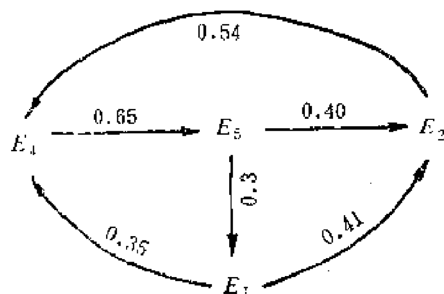


图2-8-1 沉积旋回模式图

如果加以简化，可以得到主要旋回模式为

$E_4 \to E_5 \to E_2 \to E_4$ 或

$E_4 \to E_5 \to E_1 \to E_2 \to E_4$。

第三篇　石油资源定量评价

石油是当今世界最重要的能源之一，我国自1978年原油产量超过100Mt以来，石油产量逐年稳步增长，并已进入世界主要产油国行列。我国陆地上沉积岩分布面积达5Mkm²以上，各种类型的沉积盆地200多个，海域水深在200m以内的大陆架地区有1.3Mkm²以上，这些是发展我国石油工业的雄厚物质基础。

本世纪70年代以来，由于电子计算机技术的普遍推广，数学地质方法的应用，极大地推动了我国石油资源定量评价工作的开展，并且提高了评价方法的科学性及评价结果的可靠性。

由于准确的石油资源评价结果是发展石油工业的基本依据，并且会带来巨大的经济效益，因而石油资源定量评价工作受到世界上各产油国的普遍重视。最近一段时期内，各主要产油国都定期开展区域性的石油资源定量评价工作。例如，美国、苏联等国家大体上五年左右进行一次全国性的石油资源评价。

第一章　石油资源定量评价的有关问题

石油资源定量评价是贯穿整个石油勘探过程的一项综合性预测工作。评价人员应当收集一切可以利用的地质资料，采用合理的预测方法，估算评价地区的石油、天然气蕴藏数量；预测勘探地区中的有利勘探地带；从经济技术角度进行勘探方案的可行性论证，为制定合理的勘探部署方案提供依据。

勘探方案实施后，评价人员还要根据勘探过程中获得的信息反馈，不断地修改、补充原有的勘探方案。这种由信息反馈而进行的勘探方案调整工作，要一直进行到下一轮资源评价开始时为止。所以，石油资源定量评价工作是一项既有评价的阶段性，又有评价的连续性的滚动式的动态预测过程。

第一节　石油储量与石油资源

通过地质勘探及油田开发，按照人们对石油及天然气在地壳中贮存状态的认识程度，以及开采它们的经济技术条件，可将石油及天然气在地壳中的赋存量分为两大类，即石油储量与石油资源量。

石油储量是指已经探明或基本探明的，在目前经济技术条件下，可以开采利用的那部分石油和天然气的数量；而石油资源量是指有待发现或者虽已发现，但在目前经济技术条件下还不能开采利用的那部分石油和天然气的数量。

石油储量有地质储量和可采储量之分。地质储量是指在地层埋藏条件下，在具有产出油气能力的储集层中的油气数量，石油是以地面条件下的重量单位表示；天然气是以标准状态（温度为20°C，压力为760mmHg）下的体积单位表示。而可采储量是指在目前经济技术条件下，从储集层中所能采出地面的那部分油气数量。

一、地质储量的分级

根据勘探阶段或开发阶段对含油气地质条件的认识程度，一般可将地质储量分为三个级别，即探明储量、控制储量和估算储量。

1. 探明储量

探明储量也称证实储量，是指油气藏的地质特征完全探明或基本探明，在目前经济技术条件下可以立即进行开采的准确或较准确的储量。探明储量是勘探阶段的最终成果，是油田开发建设投资的决策依据。

2. 控制储量

控制储量也称概算储量，是指油气藏地质特征尚未完全探明，含油边界还带有推测因素的储量。储量的计算参数是根据数量不足的资料所确定的，或者是通过邻区对比所确定的。控制储量可以作为制定进一步勘探计划的依据。

3. 估算储量

估算储量也称预测储量，是在经过地震勘探或其他勘探方法所确定的地质圈闭上，钻获工业性油气流后按圈闭法估算的储量。已获得工业性油气流地区的邻近圈闭，如果含油气地质条件基本相同，按圈闭法预测的储量也可作为估算储量。这些地质圈闭内的油层变化、油水关系均未搞清，含油面积、油层厚度等关键参数可能有很大出入。估算储量可作为制定评价性勘探方案的依据。

二、石油资源量的分级

石油资源量按探区勘探程度可分为三个级别，即潜在资源量、推测资源量和尚未研究的资源量。

1. 潜在资源量

潜在资源量也称圈闭性远景资源量，是指经过地质、地球物理勘探的地区用地质类比法求得的资源量。即在已有含油气综合评价资料的地区，对具有含油气远景的各类地质圈闭或构造带逐个进行地质类比统计，根据获得的半定量地质参数，用概率统计法求得的资源量的范围值。潜在资源量可以作为制定探区预探方案的依据。

2. 推测资源量

推测资源量也称类比法远景资源量，是指根据地质资料与邻区同类型沉积盆地进行对比，结合盆地或凹陷的少量物探资料、基准井或参数井的储集层物性和生油岩的有机地球化学资料，由概率统计法计算得到的资源量范围值。由盆地模拟法计算得到的资源量也属于推测资源量。推测资源量可作为编制早期区域勘探部署的长远规划的依据。

3. 尚未研究的资源量

尚未研究的资源量的含义是明确的，石油勘探空白地区的石油资源都属此类。

各级石油储量与各级石油资源量之间是一种动态结构关系，见图3-1-1。通过勘探，石

油资源量可以升级为石油储量，低级石油储量可以升级为高级石油储量。

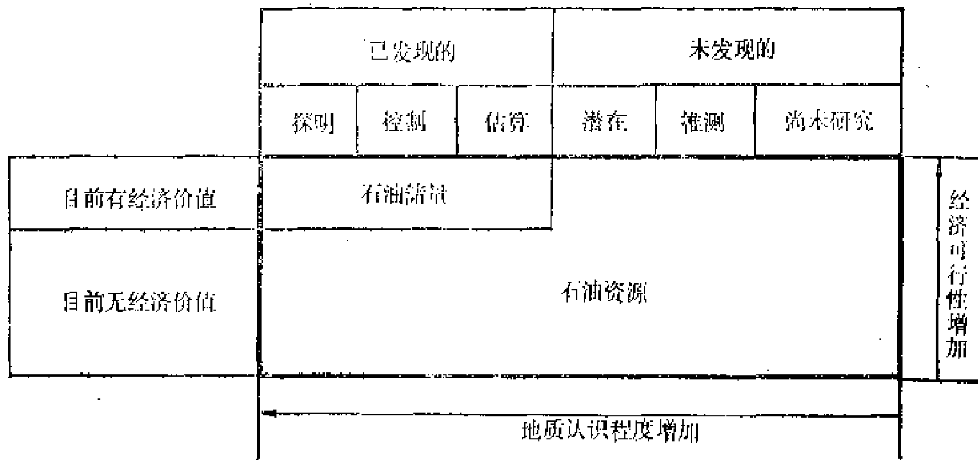


图3-1-1　石油储量与石油资源的结构关系

第二节　石油资源定量评价的任务

石油资源定量评价的任务是对某一勘探地区（例如一个地质凹陷、一个地质坳陷、一个沉积盆地、一个行政省，一个国家乃至全世界）所蕴藏的石油和天然气的数量作出估算；对石油和天然气富集地带的位置作出预测；以及对进一步勘探开发这些油气资源的技术进行可行性论证和经济效益的概算。

目前在我国，对于一个具体的勘探地区来说，石油资源定量评价主要应该回答如下三个问题：

（1）这一地区有没有石油资源？有多少石油资源？

（2）这一地区的油气富集地带在什么地方？应当如何进行勘探？

（3）这一地区进行石油勘探是否有经济效益？应当如何进行勘探部署？

如果认为石油资源评价仅仅是估算一下石油资源量，那是片面的、不准确的。而其全面、准确的含义应当包括上述三个方面。从这三个方面的内容看，石油资源定量评价工作是一项贯穿整个勘探过程的综合性预测工作。因此，从学科上讲，它应当属于预测学范畴。

有人认为，预测是科学加艺术，是客观资料与主观判断的综合，因而对同一问题的预测可采用多种不同的方法。特别是在早期或中期勘探阶段进行石油资源定量评价时，绝不要局限于一种思路、一种理论、一种模式、一种方法，而要利用一切可以利用的地质资料，从不同角度进行预测，以利于互相验证补充，以期得出尽可能接近客观实际的评价。

第三节　石油资源评价的工作要点

石油资源评价工作的要点可以概括为："从五个途径入手，抓好四个环节，发展三种软件，解决两个问题，实现一个目标"。

1. 关于五个途径

纵观国内外油气资源评价工作的现状以及评价方法的历史演变，开展石油资源评价工作都是从五个途径入手。这五个途径就是类比预测途径、外推预测途径、统计预测途径、成因预测途径、综合预测途径。

类比预测及外推预测是石油地质学中的传统预测方法。统计预测是本世纪60年代末逐渐发展起来的一类预测方法，也是目前国内外油气资源评价的主导方法，目前仍有新的统计预测方法不断出现。成因预测是本世纪70年代末、80年代初兴起的预测方法，其特点是从油气演化历史角度进行油气资源的预测。综合预测从本世纪80年代起已转向人工智能专家系统。以上这五种预测技术的研究深度，基本上可以代表一个国家的油气资源评价工作的科学技术水平。

2. 关于四个环节

鉴于近年国外油气资源评价工作的发展现状以及我国第一轮油气资源评价工作的经验，应当积极发展计算机评价技术。其基本环节为：

(1)建立概念模型　概念模型是在深入的地质研究工作基础上建立的经验性或推理性的知识描述，包括语言模型及图式模型。

(2)建立数学模型　数学模型是概念模型量化后的模型，包括静态量化模型与动态量化模型。

(3)物理模拟验证　物理模拟是检验概念模型、数学模型是否正确的唯一手段，特别是成因预测模型在很大程度上要依赖物理模拟的实验结果。

(4)地质信息管理　地质信息管理是指与油气资源评价有关的地质数据与地质图形的管理，以及数据文件与图形文件的形成。

3. 关于三种软件

目前，国内外的油气资源评价软件系统大体上可分为三大类，即：

(1)统计预测评价系统　这种系统是以统计预测为主导方法，通常包括静态与动态两种预测模型，是以某种置信水平下估计预测问题发生的可能性，亦即可以得到“可能如此”的预测结论。

(2)成因预测评价系统　这种评价系统是以成因预测为主导方法，用差分方法或有限元法求解模型的解析解或近似解，是在某种临界标准下给出预测问题演变的近似性，亦即可以得到“几乎如此”的预测结论。

(3)人工智能专家系统　这种评价系统是以专家的实践经验为依据，以知识推理方式回答问题，是在某种知识可靠条件下回答预测问题出现的合理性，亦即能够得到“应当如此”的预测结论。

4. 关于两个问题

一个大型的实用性的应用软件预测系统，必须认真解决好如下两个问题：

(1)评价系统的先进性与实用性　系统的先进性决定于地质研究工作的深入程度、概念模型与量化模型的可靠性，以及软件的优化程度；系统的实用性是指它应能最大限度地适用于不同地质条件、不同勘探时期、不同勘探对象的资源评价。

(2)评价系统的技术推广　软件是一种技术手段，不能推广的软件不可能为全国性的资源评价服务。

5. 关于一个目标

这个目标就是能以最小的人力、物力，在最短时间内，作出科学的、准确的评价结果。同时，也要考虑到软件系统要兼顾探区日常生产、科研上的需要。

第四节　预测过程的基本概念

石油资源定量评价的实质，是评价人员通过研究分析，把已获得的地质信息彼此联系起来，并把这些信息转化成为探区中石油和天然气的数量多少和空间分布位置的概念。这个转化过程包括两个基本阶段；第一阶段是由已获得的地质信息建立预测模型，即建模阶段；第二阶段是用建立的预测模型进行探区的油气资源预测，即解模阶段。

一、地质信息与地质模型

什么是信息？目前的定义很多。通常把信息理解为"可以在人们之间传播的，能使消息中所描述事件出现的不确定性减少的那些因素"。产生信息的地方称作信源。信息可以分为没有经过加工的基础信息和经过介体加工的复合信息。介体可以是主观观点或客观工具。基础地质信息是指通过一切勘探手段所观测到的地质数据，图形或资料；复合地质信息是指由基础地质信息经过主观观点（如某种地质理论、地质规律、地质认识、地质假说等）或客观工具（如仪器、计算机等）加工后生成的复合地质数据、加工图形或地质观点等。

什么是体系？一个体系可看作是从客观世界中被主观选取的一个局部。那么，在石油资源评价时，被唯一圈定的勘探地区，例如一个沉积盆地、一个行政省、一个海域招标区等等都可称为地质体系。为使一个地质体系能被地质家们认识，就得建立一个人们都能理解的地质模型。前已述及，地质模型就是对地质体系的一个表示或体现。也就是说，地质模型应当是对地质体系的一个简化或概括。因此，建立模型时需要把握如下要点：

（1）模型要有足够的精度，必须把体系的本质因素包括进去。在不影响精度的条件下，尽量排除较次要的因素。

（2）模型既要精确又要简单。如果简单的模型就能使问题得到满意的答案，则不必去搞复杂的模型。模型太复杂则失去了建立模型的本意。

（3）模型必须通过大量验证，说明是行之有效的。

（4）模型必须具有预测能力，否则所建立的模型毫无意义。

（5）模型的数学表达式应当尽量向标准的数学公式靠拢。

二、概念模型与数学模型

在对地质体系及其地质信息深刻理解的基础上，用定性的方法（包括文字叙述、图象表示）描述地质变量之间关系的模型叫概念模型或定性模型。而用定量方法以字母、数字、数学符号建立的等式、不等式、图象来描述地质变量之间关系的模型叫数学模型或定量模型。概念模型是建立数学模型的基础；由概念模型过渡到数学模型是对地质体系在认识上的深化。

为了鉴别概念模型或数学模型的可靠性，常以实验手段进行验证，例如用水槽实验模拟沉积过程，用泥巴实验、光弹实验模拟构造演变，这些实验一般称作物理模拟或物理模型。

三、经验模型与成因模型

经验模型是根据石油地质勘探及油气田开发过程中积累的大量资料总结出来的模型，经过多次验证后，可以用来表示某些地质变量之间的定量关系。经验模型最主要的特点是在数学表达式中至少有一个所谓的经验系数。经验系数的地质意义是明确的，但是对影响经验系数取值的机制并不清楚。

成因模型是根据石油地质学的基本理论推导出来的模型，经过多次验证后，可以用来表示某些地质变量之间的定量关系。成因模型的特点是在数学表达式中不允许有经验系数。

由于地质体系的复杂或者获取的信息数量的不足，经常在一些地质模型中既有经验模型的成分也有成因模型的成分，这类模型可称作复合模型。

四、确定型模型与随机型模型

用地质模型对地质体系进行预测时有个精度问题。由地质模型X得到一个预测值$\hat{X}$，它与地质体系的实际观测值Y一般并不相同。

在确定型模型中，$\hat{X}$是模型给出的一个预测值，如果与实际观测值Y的差值（$Y-\hat{X}$）是由观测误差引起的，而且大得使人不能接受，则可以舍弃Y而重新观测；如果差值（$Y-\hat{X}$）是由模型对体系的偏差引起的，而且大得使人不能接受，则可以舍弃模型X而重建模型。在决定取舍观测值Y或模型X之前，可以主观确定一个临界标准，这个标准在线性规划中叫作约束条件，在最优化过程中称可行域边界。如果以ε表示可行临界值，$|Y-\hat{X}|\leqslant\varepsilon$称为可行变程条件，它说明模型对体系的近似程度。如果实际观测值Y和模型给出的预测值$\hat{X}$的近似程度合乎要求时，则通过确定型模型可以得到“几乎如此”的预测结论。

在随机型模型中，$\hat{X}$是模型给出预测值的随机分布的数学期望。从统计观点看，模型X是随机体系的一个体现，所以对应于$\hat{X}$的实际观测值Y也应该是体系的随机分布的数学期望。而这两个随机分布未必相同，对模型X或观测值Y的取舍可用事先确定的置信水平作为临界标准。如果以$|Y-\hat{X}|\leqslant\varepsilon$来表示置信区间条件，相应的置信水平（$1-\alpha$）可以说明模型体现体系的可能性。如果随机型模型给出的预测值$\hat{X}$体现实际观测值Y的可能性合乎要求，则通过随机型模型可以得到“可能如此”的预测结论。

五、预测过程与信息反馈

地质模型X的建立是通过介体K对地质体系Y的作用O而完成的。这个过程可用关系式表示为

$$YOK=X$$

这里O可理解为广义的作用，不要单纯理解为一般算子。例如，某个地质构造为体系Y，通过地震勘探K的实施O，可以得到地震磁带上的数字序列X。而X就是这个地质构造的模型。又如，某个探区的石油资源量为体系Y，通过评价人员K对实际资料的统计分析O，得到的回归方程X就是预测石油资源量的模型。

预测过程包括由体系Y通过介体K_1的作用O_1映射得到模型X，即建模阶段；以及再以模型X作为信源，通过介体K_2的作用O_2映射得到预测结论$\hat{X}$，即解模阶段。这两个阶段可以表示为

$$YO_1K_1=X$$
$$XO_2K_2=\hat{X}$$

体系Y、模型X、预测结论$\hat{X}$可以用三个圆圈表示。圆圈Y与圆圈X相交的部分越大，说明模型X越能体现体系Y；圆圈X与圆圈$\hat{X}$相交的部分越大，说明预测结论$\hat{X}$越充分地利用了模型X；圆圈$\hat{X}$与圆圈Y相交的部分越大，说明预测结论$\hat{X}$越接近体系Y。

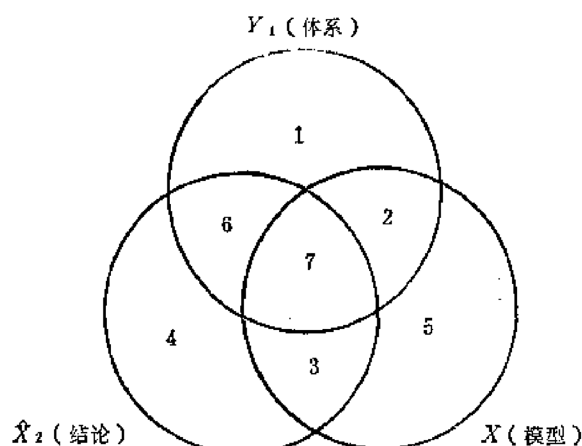


图3-1-2 预测过程的7种结果

预测过程中，如果三个圆圈互不相交时，说明模型是错误的，预测结论也必然是错误的。而只有当三个圆圈完全重合时，预测结论才完全是正确的。一般情况下，三个圆圈之间互相相交可形成7个区。这七个区表示7种预测结果，见图3-1-2及表3-1-1。

1区：建模中漏失的信息，称一次一型错误。

2区：解模中漏失的信息。称二次一型错误。

3区：建模中引入的假信息，称一次二型错误。

4区：解模中引入的假信息，称二次二型错误。

5区：由假信息建立的错误模型，并且没有进行预测。当然，不预测不会出错，但是，应当预测而不预测本身就是一种错误。

6区：没有利用模型进行预测，却又碰巧得出正确结论，所以是偶然正确。偶然正确不

表3-1-1 预测过程的分区说明

区号	Y	X	$\hat{X}$	预测结论的性质	
1	+			一次一型错误	漏失信息
2	+	+		二次一型错误	
3		+	+	一次二型错误	引入假信息
4			+	二次二型错误	
5		+		应当预测而没有预测	
6	+		+	没有利用模型进行预测	
7	+	+	+	非完全正确	

能扩张，被引用时必然产生错误。

7区：表示用正确的模型得出正确的预测结论。但是，并非是完全正确，只有当三个圆圈完全重合情况下，才是完全正确。

利用石油资源评价得出的预测结论，可以指导探区的勘探工作，在这一过程中，又可以从地质体系获得新的地质信息，这些信息既可用以检验原来的预测结论，也可用来改进预测模型，近而由改进的预测模型得到新的预测结论，使其更加接近实际。这种过程，也就是信息反馈的过程的反复进行，显然可以逐步提高预测结论的可靠性。因而，石油资源定量评价是一个信息反馈的动态预测过程。

第五节　石油资源评价的理论基础

石油资源定量评价使用的各种模型都是根据某种理论建立的，这些理论也是石油资源定量评价方法的分类原则之一。

一、统计预测原理

地质学所研究的地质现象和地质过程都普遍地受概率法则支配，也就是说地质现象和地质过程可视为随机事件，因而由各种观测手段得到的大多数地质信息都具有随机变量性质。从统计学观点看，石油地质勘探过程的实质就是对所研究的地质体系的抽样观测过程。这里把直接观测到的地质数据看作基础地质信息，而把由统计方法得到的统计量看作复合地质信息。随着勘探工作的不断进行，积累的资料逐步增加，才有可能得到对地质体系的无偏估计，也就是说，才有条件对所研究的问题得出“可能如此”的预测结论。这就是石油资源定量评价的统计预测原理。由统计预测原理建立的预测模型，显然都属于随机型模型。

在以往预测石油、天然气数量时，大都采用参数法，其中最常用的是与体积参数有关的参数法。这些参数法在算法上可归结为一些地质参数的连乘，而每个参数常以随机变量的平均值作为代表值参加运算，最后得到唯一的一个资源量预测值，而对这个预测值的可靠性如何并不清楚。近年来，由于使用了统计预测方法，例如蒙特卡罗法，则可给出不同概率水平下的资源量估计值；又如使用了各种多元统计分析方法，不仅可以预测石油、天然气的数量，而且可以预测探区中的含油气有利勘探地带。

二、外推预测原理

外推预测就是根据体系自身已经确知的变化规律，经过延伸去预测地质体系的未来演变过程。任何一种地质体系，如果把其看作一个有序集合，则序集的外延就是外推预测。也就是说，通过认真分析探区中已经得到的实际资料，可以建立拟合模型去逼近以往的勘探或开发历程，当拟合精度达到要求时，这个模型的外延部分就可以表示这个探区未来的勘探或开发的前景，这就是外推预测原理。其中时间序列的外延是最常用的外推预测方法，一般称作时间序列分析。由于这类模型都是根据以往的历史资料所建立的，所以，也常称作历史外推法或历史统计法。

外推预测法一般是把石油资源量、石油储量、石油产量作为时间、投入工作量或投资数额的函数，因而属于广义的时间序列分析。一般由外推预测原理建立的预测模型大多数都属

于确定型模型。近年来，在外推预测方法中已出现许多行之有效的预测方法。

三、类比预测原理

类比就是根据两个体系中已经确知的互相类似的性质，预测其尚未确知的互相类似的性质。若两个地质体系的成因相似，则其含油气地质条件也可能相似。在特定地质条件下可形成特定类型的油气藏，而在相似地质条件下则可能形成类似的油气藏。这就是石油资源定量评价的类比预测原理。由类比预测原理出发，可以根据成熟探区获得的比较全面的地质信息，建立各式各样的预测模型（包括确定型模型和随机型模型），去预测含油气地质条件相类似探区的油气资源量与油气分布规律。

成熟探区是指投入了较多的勘探工作量，对其含油气地质条件、石油和天然储量以及油气分布规律都比较清楚的探区，成熟探区也称为模型区或实习区。被预测的地区也称作评价区或靶区。

四、成因预测原理

石油、天然气在地壳中的生成、运移、聚集直至形成油气藏，是各种控制油气形成的地质因素在地质历史中演化、搭配的结果。生油条件、储油条件，圈闭条件、运移条件、聚集条件、保存条件等等都是控制油气藏形成的最基本的地质因素。

近年来，在研究油气生成、运移、聚集等油气藏形成机理方面已取得很大进展。尽管这些研究结果尚待进一步深化，但不可否认，这些理论或假说正在指导或影响着石油地质勘探工作。按着石油，天然气的成因机理或假说预测探区中的油气资源数量，探索油气分布规律，就是石油资源定量评价的成因预测原理。由成因预测原理导出的预测模型，原则上都应属于确定型模型。但是，由于目前成因理论上的不完善，往往在成因模型中仍有个别的经验系数，也就是说，还有某些经验模型的成分。

石油、天然气在地壳中的形成和演化是个极其复杂的物理化学过程，涉及到古地温、古压力、界面效应、滤流过程等等物理演化，以及有机物质化学演化的一系列过程。所以，成因预测方法的技术难点相当多，其中的许多问题都涉及到石油地质学的基础理论。但是，成因预测方法的深入研究，不仅利于石油资源定量评价，并且可以从根本上改变石油地质学中一些传统的陈旧观念，从而丰富和发展石油地质学的基础理论。

五、综合预测原理

任何一个地质问题都具有时间长、空间广、因素多的特点，因而使问题变得十分复杂。同时，由于石油地质勘探阶段，特别是早期勘探阶段所能了解到的情况又很不全面，这就容易使预测结论犯“弃真”或“取伪”错误，例如，在含油气地区漏算储量或漏圈油气藏；或在不含油气地区错算了储量或错圈了油气藏。

实际上通过地质勘探得到的一切资料，都应看作是从不同侧面向评价人员提供的找油信息。为了克服认识上的局限性，则应该利用一切可以利用的找油信息，进行信息的加工与综合，以便得出尽可能全面的预测结论。这就是石油资源定量评价的综合预测原理，例如特尔菲法就是典型的综合预测方法。由于特尔菲法综合了多个专家的评价意见，所以对信息保真也是有益的。又如多种信息迭合评价法也是一种尽最大可能地利用各种地质信息来预测探区

中有利勘探地带的方法。

近年来，计算机人工智能已得到迅速发展，并以专家系统的方式用于石油资源评价。目前所谓的专家系统就是用人工智能的理论和方法，建立的一个具有石油地质专家技术水平的计算机应用软件系统。从评价理论分类上看，它应属于综合预测原理。

第六节 评价方法的分类原则

据不完全统计，目前国内外已有的各种具体的石油资源定量评价方法多达百余种，常用的方法也有二、三十种之多。对这些众多的评价方法进行合理分类显然是很必要的。但是，合理分类并不容易，因为按任何原则进行分类都不能避免某些方法之间的相互交叉，有些方法既有确定型模型成分，又有随机型模型成分；或者既有经验模型成分，又有成因模型成分。下面给出四种评价方法的分类原则。

1. 按石油资源评价任务，可将评价方法分为三大类：

（1）石油、天然气的资源量预测方法；

（2）含油气有利地带的预测方法；

（3）油气勘探的经济评价与决策分析方法。

2. 按石油资源评价的理论基础，可将评价方法分为五大类：

（1）统计预测方法；

（2）外推预测方法；

（3）类比预测方法；

（4）成因预测方法；

（5）综合预测方法。

3. 按油气勘探开发阶段，可将评价方法分为四大类：

（1）早期勘探阶段预测方法；

（2）中期勘探阶段预测方法；

（3）晚期勘探阶段预测方法；

（4）油田开发阶段预测方法。

4. 按预测模型的性质，可将评价方法分为两大类：

（1）确定型预测方法；

（2）随机型预测方法。

下面按石油资源评价的任务、理论基础、勘探开发阶段，将常用的三十多种评价方法列于表3-1-2中。

表3-1-2　石油资源评价方法的分类

评价方法分类 \ 勘探开发阶段	早期 勘探阶段	中期 勘探阶段	晚期 勘探阶段	油田 开发阶段
一、石油资源量预测方法				
1. 统计预测方法				
（1）蒙特卡罗法	—	—	—	
（2）回归分析方法	—	—	—	—
2. 外推预测方法				
（1）指数函数模型	—	—	—	
（2）逻辑斯特模型	—	—	—	—
（3）Weng旋回模型				—
（4）大油田与中小油田比例模型	—			
（5）储量变化率与增长率模型	—	—		
（6）油田规模序列法	—	—	—	
（7）油藏规模分布法	—	—		
3. 类比预测方法				
（1）储量密度系数法	—			
（2）聚集系数法	—			
（3）储集层体积法	—			
（4）单位产率法	—			
（5）评分法	—			
（6）地质因素比较法	—			
（7）断层线密度法	—			
（8）构造平均法	—			
（9）容积系数法	—			
（10）地质因素集合取小法	—			
（11）沉积速度法	—			
（12）油田模型法	—	—		
4. 成因预测方法				
（1）剩余沥青法	—			
（2）运移系数法	—			
（3）埃德曼法	—			
（4）干酪根降解模型法		—	—	
（5）盆地动态模拟法		—	—	
5. 综合预测方法				
（1）特尔菲法	—	—	—	
（2）人工智能专家系统评价方法	—	—	—	
二、含油气有利地带预测方法				
1. 统计预测方法				
（1）回归分析方法	—	—	—	—
2. 外推预测方法				
（1）趋势分析方法	—	—	—	

续表

勘探开发阶段 评价方法分类	早期 勘探阶段	中期 勘探阶段	晚期 勘探阶段	油田 开发阶段
3. 类比预测方法				
（1）聚类分析方法	——	——	——	——
（2）判别分析方法	——	——	——	——
（3）因子分析方法	——	——	——	——
（4）对应分析方法	——	——	——	——
4. 成因预测方法				
（1）盆地动态模拟法		——	——	
5. 综合预测方法				
（1）多种信息综合评价法	——	——		
（2）模糊集合综合评价法	——	——		
（3）人工智能专家系统评价方法	——	——	——	——
三、经济评价与决策分析方法				
1. 最优化方法				
（1）线性规划方法	——	——	——	——
（2）动态规划方法	——	——	——	——
2. 决策方法				
（1）最大可能法	——	——		
（2）期望值法	——	——		
（3）决策树法	——	——		

第二章　预测石油资源量的主要方法

前已述及，目前国内外已有的石油资源量预测方法多达百余种，本书不能一一介绍，本章只讲述一些具有代表性的石油资源量预测方法。

第一节　蒙特卡罗法

蒙特卡罗（Monte—Carlo）法亦称统计模拟法，50年代的译文中也曾称其为统计试验法。该方法的实际应用和系统发展始于本世纪40年代。但如果从问题的提出开始，却可以追溯到十七世纪后半叶，法国著名学者布丰（Buffon）1777年通过随机投针试验，发现了随机投针的概率与圆周率π之间的关系。根据布丰给出的用投针试验求π的近似公式，不少人曾作过几千至几十万次投针试验，得出π的估计值为3.14至3.14159。

本世纪40年代电子计算机的出现，才有可能实现大量的随机抽样试验，并用蒙特卡罗法来解决实际问题。当时最有代表性的实际应用，是在第二次世界大战期间用于原子弹的研制方面，具体地说就是在电子计算机上对中子的行为进行随机抽样模拟，来推断所要求的参数。1946年，物理学家冯·诺曼（Von Neumann）用随机抽样方法模拟了中子连锁反应，当时出于保密而将这种方法以赌城蒙特卡罗命名为蒙特卡罗法。

蒙特卡罗法的现代含义是利用各种不同分布的随机变量抽样序列，模拟给定问题的概率统计模型，给出问题数值解的渐近统计估计值。或者简要地说，蒙特卡罗法是应用随机数技术进行模拟计算的方法的统称。目前，蒙特卡罗法的实际应用大体上包括如下四个方面：

（1）对给定问题建立简化的概率统计模型，使所求得的解恰好是所建立模型的概率分布或者数学期望；

（2）研究生成伪随机数的方法以及研究各种实际分布产生随机变量的抽样方法；

（3）根据统计模型的特点和实际计算的要求，进一步改善模型，使之降低方差和提高计算效率；

（4）给出获得求解问题的统计估计值以及方差或标准误差的方法。

从以上解决问题的方面看，蒙特卡罗法是一种应用领域非常广泛的通用性的统计学方法，而并不是石油资源定评价特有的方法。从应用的内容看，蒙特卡罗法用于石油资源定量评价主要是第二个方面，即研究生成伪随机数的方法以及研究与石油资源有关的随机变量的抽样方法等问题。

蒙特卡罗法用于石油资源定量评价开始于本世纪60年代。美国于1975年完成的第二次全美石油资源评价的主要算法就是蒙特卡罗法。目前世界各主要产油国及西方各大石油公司都把蒙特卡罗法作为石油资源定量评价的重要方法之一，广泛应用于含油气地区的早、中期勘探阶段。我国应用蒙特卡罗法估算石油资源量开始于1979年，现在国内各油田已普遍使用，并已成为以统计预测为主的应用软件评价系统的核心算法。

石油勘探阶段，特别是在早期勘探阶段，较准确地估算勘探地区的石油资源量是十分重

要的，因为其后的石油勘探、油田开发以及油田建设的决策，完全取决于这一地区石油资源的数量。诚然，一个勘探地区有没有石油资源，有多少石油资源，完全由这个地区的含油气地质条件所决定。但是，在勘探阶段由于勘探人员对研究地区的含油气地质条件在认识上的不完全，致使未来的勘探成效具有很大的不确定性，国外把这种勘探成效的不确定性称作石油勘探的风险。因而，勘探人员特别需要用概率统计方法去处理分析勘探过程中已掌握的地质资料与石油资源之间的内在联系，让主观臆想成分尽可能减少，使勘探工作立足于最现实的可能性上，以利于提高勘探成效。

勘探人员基于不同的找油理论，对一个勘探地区石油资源量的估算可有不同的方法，但是，含油气区中任何一个局部含油地质单元（例如局部构造、断块、单斜、岩性圈闭等）的石油资源量最常用的计算公式，都可以归结为一些地质参数与经验系数的连乘，即

$$Q_j=\prod_{i=1}^{n}X_{ij} \qquad (j=1,2,\cdots,m) \tag{3-2-1}$$

式中 Q_j——含油区中第j个局部含油地质单元的石油资源量；

X_{ij}——第j个局部含油地质单元的第i个地质参数或经验系数。

而一个含油区石油资源总量的计算公式，可表示为全部的m个局部含油地质单元的石油资源量的累加，即

$$Q=\sum_{j=1}^{m}Q_j=\sum_{j=1}^{m}\prod_{i=1}^{n}X_{ij} \tag{3-2-2}$$

式中 Q——含油区的石油资源总量。

（3-2-2）式中的X_{ij}可以是性质不同的地质参数，一般有两种可能，其一是随机变量；其二是常数或经验系数。如果是随机变量，则需要在计算石油资源量之前构造其分布函数。

用蒙特卡罗法估算一个含油区石油资源量的计算过程，大体可以归结为三个步骤，即构造随机变量的分布函数、计算局部含油地质单元的石油资源量以及计算含油区的石油资源总量。其中后两个步骤都要对一些随机变量进行抽样计算，因而研究随机数的生成与检验，就成为用蒙特卡罗法估算石油资源量的一个重要技术环节。

一、随机数的产生和检验

用蒙特卡罗法模拟一个实际工程技术问题时，要用到数以千计、万计、甚至百万计的随机数。因此，在模拟计算之前必须成功地解决生成符合要求的随机数的问题。或者说在计算机上快速、经济地产生各种不同分布的随机数是蒙特卡罗法能够成功应用的基础。

所谓随机数就是各种不同分布随机变量的抽样序列，在序列内部无任何规律可寻。日语中随机数的汉语意义是"乱数"，就是说序列的排列是杂乱无章的或者是完全随机的。模拟石油资源量时，经常用到均匀分布的随机数，有时也用到正态分布的随机数。

早些年人们曾把事先造好的随机数表存入计算机中，使用时随时调用。也有人将放射性物质的随机放射源或物理噪声的随机噪声源与计算机联接，把随机的物理过程转变为随机数。目前这两种方法已不再使用。

随机数表包含从0到9十个数字，其排列是完全随机的，表中的任意一位上的数字为p（$p=0,1,2,\cdots,9$）的概率都等于1/10，而且与上下左右相邻的其他数字的出现都是不相关

的。满足这些条件的数字就是由0到9十个离散的随机数字。如果把相邻的四个数字合并，再除以10^4，则可把这些数字看成是0.0000到0.9999之间均匀分布的随机数。

目前，发展最快且使用最广泛的是用数学方法产生随机数。但是，严格地讲用数学方法根本不能产生真正的随机数，因而通常把用数学方法产生的随机数称作"伪随机数"。尽管如此，只要对伪随机数序列进行一系列严格的统计检验，证明其可以满足模拟计算的精度要求，则伪随机数就可以作为真正的随机数来使用。

为了满足模拟问题的实际需要，要求在计算机上产生随机数的速度要快，占用计算机的内存要小，产生的随机数序列要有足够长的周期，而且应当具有符合要求的概率统计性质。

从理论上讲，只要有一种连续分布的随机数，就可以通过数学变换产生其他分布的随机数。在连续分布函数中，〔0，1〕闭区间上标准均匀分布的随机变量是最简单、最基本的一种。因而习惯上有人把〔0，1〕上均匀分布的随机变量的抽样值就称作随机数。而其他分布的随机数都可以借助均匀分布随机数来产生，所以说均匀分布随机数是随机抽样的基本工具。

设R是〔0，1〕上的标准均匀分布的随机变量，则它的密度函数为

$$f(x)=\begin{cases}1 & (0\leqslant x\leqslant 1)\\ 0 & (x<0, x>1)\end{cases} \tag{3-2-3}$$

它的分布函数为

$$F(x)=\begin{cases}0 & (x<0)\\ x & (0\leqslant x\leqslant 1)\\ 1 & (x>1)\end{cases} \tag{3-2-4}$$

R的数学期望为

$$E(R)=\int_0^1 xf(x)\mathrm{d}x=\int_0^1 x\mathrm{d}x=\frac{1}{2}$$

方差为

$$D(R^2)=\int_0^1 (x-\frac{1}{2})^2\mathrm{d}x=\frac{1}{12}\approx 0.08333$$

1. 随机数的产生

在计算机上用数学方法产生随机数的方法有迭代取中法、移位法、同余法。

（1）迭代取中法　平方取中法是迭代取中法的一种基本算法，是用数学方法产生伪随机数的最早方法，也曾经是Neumann所提倡的方法。

把一个b进制$2k$位的数X_0作为种子数自乘后，一般可以得到一个$4k$位的数X_0^2，它们可以分别记为

$$X_0=b_{2k}b_{2k-1}\cdots b_2b_1$$
$$X_0^2=c_{4k}c_{4k-1}\cdots c_2c_1$$

把X_0^2截头去尾，取中间$2k$位数字，即取从$k+1$位数字至$3k$位数字作为X_1，即

$$X_1=c_{3k}c_{3k-1}\cdots c_{k+2}c_{k+1}$$

然后除以b^{2k}，作为〔0，1〕上的第一个伪随机数。对X_1重复上述过程，即将X_1^2截头去尾留中间$2k$位数字作为X_2，除以b^{2k}作为〔0，1〕上的第二个伪随机数。如此反复进行这一过程，直至出现的伪随机数退化为0，或者出现的伪随机数与前面的随机数重复，即出现了周

期性时为止。

对于10进制即为

$$\begin{cases} X_{n+1} \equiv \left[\dfrac{X_n^2}{10^k}\right] & (\text{mod } 10^{2k}) \\ \gamma_{n+1} = X_{n+1}/10^{2k} \end{cases} \quad (3\text{-}2\text{-}5)$$

对于2进制则为

$$\begin{cases} X_{n+1} \equiv \left[\dfrac{X_n^2}{2^k}\right] & (\text{mod } 2^{2k}) \\ \gamma_{n+1} = X_{n+1}/2^{2k} \end{cases} \quad (3\text{-}2\text{-}6)$$

写成b进制的通式为

$$\begin{cases} X_{n+1} \equiv \left[\dfrac{X_n^2}{b^k}\right] & (\text{mod } b^{2k}) \\ \gamma_{n+1} = X_{n+1}/b^{2k} \end{cases} \quad (3\text{-}2\text{-}7)$$

（3-2-7）式中

X_n ——第n步的随机数；

X_{n+1}——第$n+1$步的随机数；

$\left[\dfrac{X_n^2}{b^k}\right]$——不超过实数$\dfrac{X_n^2}{b^k}$的最大整数部分，即对$\dfrac{X_n^2}{b^k}$取整。

因而，（3-2-7）式的含义是整数$\left[\dfrac{X_n^2}{b^k}\right]$除以模$M$后的余数，数论上称之为以$M$为模的同余式。

例如，取种子数$X_0=6406$，10进制的2k=4，则有

$$X_0^2 = 41036836$$

$$\left[\frac{X_0^2}{10^2}\right] = 410368$$

$$X_1 \equiv 410368 \quad (\text{mod } 10^4)$$

即为$410368/10^4$的余数，所以

$$X_1 = 0368$$

如此反复，则有如下序列，见表3-2-1。

由表3-2-1可见，$X_{20}=6100$与$X_{16}=6100$完全相等，即开始出现重复。也就是说，用这一随机序列只能生成20个〔0，1〕上的随机数。

此外，用平方取中法产生的随机数，如果中间$2k$位数字中最前面的位数$c_{3k}c_{3k-1}\cdots$为0时，则下一步自乘时便得不到$4k$位数字，这样就使随机序列逐步变小，最后退化为0。

这两种情况使得平方取中法产生的随机数序列不易稳定，周期过短。所以，目前平方取中法已很少使用。Metropolis对平方取中法作过深入研究，取2进制，$2k=38$时，用平方取中法可产生一些较长的序列，最长的接近75万次，最后退化为0。然而，75万个与2^{38}相比毕竟太短，而且退化现象不可避免。

对平方取中法进行改进后可以得到各种类型的迭代取中法。例如，乘积取中法就是其中

表3-2-1　平方取中法的随机序列

序　号	X_n	X_n^2	序　号	X_n	X_n^2
0	6406	41036836	11	0609	00370881
1	0368	00135424	12	3708	13749264
2	1354	01833316	13	7492	56130064
3	8333	69438889	14	1300	01690000
4	4388	19254544	15	6900	47610000
5	2545	06477025	16	6100	37210000
6	4770	22752900	17	2100	04410000
7	7529	56685841	18	4100	16810000
8	6858	47032164	19	8100	65610000
9	0321	00103041	20	6100	
10	1030	01060900			

的一种，这种方法是任意取两个种子数X_0及X_1作为初值，即

$$\begin{cases} X_{n+2}\equiv\left[\dfrac{X_nX_{n+1}}{b^k}\right] \quad (\mathrm{mod}\ b^{2k}) \\ \gamma_{n+2}=X_{n+2}/b^{2k} \end{cases} \qquad (3\text{-}2\text{-}8)$$

γ_{n+2}就是〔0，1〕上的伪随机数。与平方取中法相比，用乘积取中法产生的伪随机数，在其均匀分布性以及序列周期长度方面都有所改进。但是，效果也不太理想。

（2）移位法　移位法产生伪随机数的作法，是在字长$2k$位的b进制的计算机上，取一个初值X_0，将X_0左移p位得到X_{01}，右移p位得到X_{02}，舍去X_{01}与X_{02}左右两边超出的位数，然后相加得到X_1，再对X_1重复上述过程得到X_2，如此进行重复，得到序列$\{X_n\}$。最后取$\{\gamma_n=|X_n|/b^{2k}\}$作为〔0，1〕上的伪随机数序列。

移位法产生伪随机数的递推公式为

$$\begin{cases} X_{n+1}\equiv X_nb^p+[X_nb^{-p}] \quad (\mathrm{mod}\ b^{2k}) \\ \gamma_n=|X_n|/b^{2k} \end{cases} \qquad (3\text{-}2\text{-}9)$$

应当指出，计算伪随机数γ_n时，若X_n不取绝对值，则将得到〔−1，1〕上的伪随机数。

Pakob曾建议，初值可选0.142859至0.833162之间的数。徐钟济选用初值为0.833162，到出现周期性时，其序列长度约为12万次；如果选择恰当的初值，序列长度可增加到20万次左右。

Галенко介绍选用初值$X_0=1.71285$，用下面的递推公式生成伪随机数：

$$X_{n+1}\equiv X_n2^7+X_n \qquad (\mathrm{mod}\ 2^{35})$$

用移位法产生伪随机数的周期，因初值X_0的选值不同而异。一般情况下，很难得到长周期的伪随机数序列。所以，目前也很少使用。

（3）同余法　同余法是用数论中同余运算产生伪随机数的一类方法的总称，包括乘同余法、加同余法、混合同余法以及组合同余法等。目前，在计算机上用数学方法产生伪随机数大都采用同余法。

①乘同余法　乘同余法产生伪随机数的递推同余式为

$$\begin{cases} X_{n+1} \equiv aX_n \pmod{M} \\ \gamma_n = X_n M^{-1} \end{cases} \tag{3-2-10}$$

式中　X_n、X_{n+1}——分别为第n次、第$n+1$次的伪随机数；

a——乘子系数；

M——模；

γ_n——〔0，1〕上的第n个伪随机数。

乘同余法首先是由Lehmer提出，他曾取$M=10^8+1$，$a=23$，$X_0=47594118$，而得到8位10进制的伪随机序列，周期长为5 882352，并对5000个伪随机数进行了统计检验，其结果认为是满意的。

用乘同余法产生伪随机数时，种子数X_0最好选取一个$4a+1$型的数，a为任意整数；乘子系数a可选取5^{2k+1}型正整数，其中k为5^{2k+1}在计算机上所能容纳的最大奇数。

用乘同余法产生的伪随机数，得到的最大可能周期为$T=2^{k-2}$。为了使其周期较长，初值X_0、乘子系数a按下面取值是可行的，即：

乘子系数a和模$M=2^k$之间互素，并且属于由同余式

$$a \equiv \pm 3 \pmod{8}$$

给定的同余类，亦即

$$a = 8a \pm 3$$

此处a为任一正整数。而初值X_0可选取任一奇数。

②加同余法　由于加法运算速度较乘法速度为快，所以又出现了加同余法。加同余法产生伪随机数的递推公式为

$$\begin{cases} X_{n+k} \equiv X_n + X_{n+1} + \cdots + X_{n+k-1} \pmod{M} \\ \gamma_n = X_n M^{-1} \end{cases} \tag{3-2-11}$$

上式中的初值X_0，X_1，…，X_{k-1}可以是任意选定的正整数。

当$k=2$时，有

$$X_{n+2} \equiv X_n + X_{n+1} \pmod{M} \tag{3-2-12}$$

（3-2-12）式中的初值X_0，X_1可为互素的正整数。用（3-2-12）式产生的伪随机数序列中，相邻两项出现的频率较高，这说明相邻两项X_n与X_{n+1}之间不是相互独立的。因而，由（3-2-12）式产生〔0，1〕上的伪随机数，其统计性质往往不能满足要求，其随机性和独立性可能都不太好。为此，可对（3-2-12）进行改进，其中常用的方法是选择$\{X_n\}$中的部分序列，例如偶数序列，即用$\{\gamma_n = X_{2n} M^{-1}\}$产生〔0，1〕上的伪随机数。

当取模$M=2^k$时，由（3-2-12）式产生伪随机数序列的最大周期为模的1.5倍，即：

$$T = \frac{3}{2} M$$

加同余法与乘同余法之间有一定的联系，（3-2-12）式相当于乘同余式

$$X_{n+1} \equiv aX_n \pmod{M}$$

其中

$$\alpha=\frac{1+\sqrt{5}}{2}$$

$$X_0=\frac{1}{\sqrt{5}}$$

但是，它打破了乘同余法中对初值X_0、乘子系数α的一些有效约定。因此，在效果上加同余法不如乘同余法。

③混合同余法　混合同余法产生伪随机数的递推公式为

$$\begin{cases} X_{n+1}\equiv\alpha X_n+\beta \quad (\mathrm{mod}\ M) \\ \gamma_n=X_n M^{-1} \end{cases} \tag{3-2-13}$$

式中　X_n、X_{n+1}——分别为第n次、第$n+1$次的伪随机数；

α——乘子系数；

β——增量；

M——模；

γ_n——〔0，1〕上的第n个伪随机数。

当$\beta=0$时，则为乘同余法，可见乘同余法是混合同余法的特例。

用混合同余法产生伪随机数，取模$M=2^k$可以得到周期$T=2^k$的等模周期伪随机数序列，这是混合同余法优于乘同余法之处。为了得到这一最大周期，（3-2-13）式中的初值X_0、乘子系数α、增量β按下面取值是可行的，即：

乘子系数α和模$M=2^k$之间互素，并且属于由同余式

$$\alpha\equiv1 \quad (\mathrm{mod}\ 4)$$

给定的同余类，亦即

$$\alpha=4a+1$$

此处a为任一正整数。增量β为任一奇数，初值X_0为任一正整数。

对于乘同余法和混合同余法，乘子系数α起着决定伪随机数统计性质的重要作用。而混合同余法中，选择合适的增量β，也可以使伪随机数的统计性质变好。但是，选择β时必须注意其序列的相关性。

根据Rotenbergs的经验，取$M=2^{35}$，$\alpha=7$，$\beta=1$，$X_0=1$时，所得到的伪随机数序列，经过统计检验认为是合格的。

按作者的经验，用混合同余法产生伪随机数，各参数取下列数值是可行的：

$$M=2^{19}=524288$$

$$\alpha=5^5=3125$$

$$\beta=3,\ 7,\ 11,\ 17$$

$$X_0=23,\ 11,\ 19,\ 37$$

此处β与X_0分别给出四套配对数值，例如取$\beta=7$，$X_0=11$则为其中的一对数值，由于乘子系数α满足$\alpha\equiv1(\mathrm{mod}\ 4)$给定的同余类，所以生成的伪随机数序列的周期$T=2^{19}=524288$。一般情况下，周期这样长的随机数序列，已经可以满足预测石油资源量模拟计算的实际需要。而且用这些参数生成伪随机数可以在字长为16位的微型计算机上实现。

（4）同余法生成伪随机数序列的整体相关性　用乘同余法或混合同余法产生伪随机数

时，可以事先确定序列整体的自相关系数$\rho(X_n, X_{n+1})$。如果选定的参数α、β能使$\rho(X_n, X_{n+1})$趋近于0，便可以符合统计性质上的独立性要求。

对于混合同余法的（3-2-13）式，Greenberger设

$$\alpha X_n+\beta=qM+\gamma,$$

并用Dedkind倒数求和法求得伪随机数序列的整体相关系数。

当$\beta>\alpha$时，求得

$$\rho(X_n,X_{n+1})\approx\frac{1}{\alpha}+\frac{6\beta}{M}\left(1-\frac{\beta}{M}\right)+\frac{12}{M}\left(\frac{S}{\alpha^2}-\frac{\alpha}{4}\right)$$

其中　$S=\sum_{q=1}^{\alpha}q\overline{\gamma_q}$，而$\overline{\gamma_q}$是和给定$q$联系的最小$\gamma$。

当$\beta<\alpha$时，求得

$$\rho(X_n,X_{n+1})\approx\frac{1}{\alpha}+\frac{12}{M}\left(\frac{S}{\alpha^2}-\frac{\alpha}{4}\right)$$

其中　$S=\sum_{q=1}^{\alpha-1}q\overline{\gamma_q}$

由这两种情况，有如下不等式成立，即

$$-\frac{\alpha}{M}\leqslant\frac{12}{M}\left(\frac{S}{\alpha^2}-\frac{\alpha}{4}\right)\leqslant\frac{\alpha}{M}$$

从而有　　$3S\leqslant\alpha^3\leqslant 6S$。

对于乘同余法的（3-2-10）式，如果取

$$\alpha\equiv5\pmod{2^k}$$

则生成的随机数序列可以得到最大周期$T=2^{k-2}$。

若取$\alpha X_n=\gamma_n+q_nM$，其中q_n、γ_n为非负整数，由X_n唯一确定，从而可以求得随机数序列的整体相关系数。当$\alpha\leqslant2^{k-2}$时

$$\rho(X_n,\ X_{n+1})\approx\frac{1}{\alpha}+\frac{12}{M}\left(\frac{S}{\alpha^2}-\alpha\right)$$

其中　$S=\sum_{q=0}^{\alpha-1}q\overline{\gamma_q}$，而$\overline{\gamma_q}=\min\gamma_q$，若

$$\frac{3}{4}S\leqslant\alpha^3\leqslant\frac{3}{2}S$$

则有

$$-\frac{4\alpha}{M}\leqslant\frac{12}{M}\left(\frac{S}{\alpha^2}-\alpha\right)\leqslant\frac{4\alpha}{M}$$

根据上面的相关系数$\rho(X_n, X_{n+1})$公式，可以得出如下结论：

①当$\alpha\ll M$时，$\rho(X_n, X_{n+1})\approx\frac{1}{\alpha}$；

②若$\alpha=5^{2k+1}$接近M，则由统计检验表明其是一种可行的选择。

2. 随机数的检验

前已述及，用数学方法不可能产生真正的随机数。因此，对于由任何数学方法生成的伪

随机数序列，都要经过严格的检验，这种检验称作随机性检验。随机性检验包括参数检验、均匀性检验、独立性检验、组合规律性检验、无连贯性检验等各种统计检验。而每种检验中又包括多种具体检验方法，这里仅介绍其中的主要检验方法。

（1）参数检验　参数检验也称矩检验，是对伪随机数的各阶矩统计量的一种显著性检验。

所生成的n个伪随机数，各阶矩为

$$m_k = \frac{1}{n}\sum_{i=1}^{n}\gamma_i^k \tag{3-2-14}$$

式中　m_k——第k阶矩；

n——伪随机数的总数；

γ_i^k——第i个伪随机数的k次乘方。

而二阶中心矩（方差）为

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(\gamma_i - m_1)^2 = m_2 - m_1^2 \tag{3-2-15}$$

式中　m_1——一阶矩；

m_2——二阶矩。

理论上的标准均匀分布的各阶矩是已知的，r阶矩μ_r、总体方差σ^2分别为

$$\mu_r = \frac{1}{r+1} \quad ; \quad \sigma^2 = \mu_2 - \mu_1^2 = \frac{1}{12}。$$

如果生成的伪随机数符合均匀分布，则伪随机数的各阶矩m_k、二阶中心矩S^2与标准均匀分布的各阶矩μ_k、总体方差σ^2应当一致。

若以n个伪随机数为一组，总共计算了M组一阶矩m_{1i}、二阶矩m_{2i}、二阶中心矩S_i^2，并以$\overline{m_1}$、$\overline{m_2}$、$\overline{s^2}$表示它们的平均值，即

$$\overline{m_1} = \frac{1}{M}\sum_{i=1}^{M} m_{1i}$$

$$\overline{m_2} = \frac{1}{M}\sum_{i=1}^{M} m_{2i}$$

$$\overline{s^2} = \frac{1}{M}\sum_{i=1}^{M} s_i^2$$

根据中心极限定理，当$M\to\infty$时，$\overline{m_1}$、$\overline{m_2}$、$\overline{s^2}$的分布趋于标准正态分布，其平均值应分别趋于$\frac{1}{2}$、$\frac{1}{3}$、$\frac{1}{12}$，而总体方差应分别趋于$\frac{1}{12n}$、$\frac{1}{45n}$、$\frac{1}{180n}$。

因而，可建立统计量

$$U_1 = \frac{\overline{m_1} - \frac{1}{2}}{\sqrt{\frac{1}{12nM}}}$$

$$U_2=\frac{\overline{m}_2-\frac{1}{3}}{\sqrt{\frac{4}{45nM}}}$$

$$U_3=\frac{\overline{s}^2-\frac{1}{12}}{\sqrt{\frac{1}{180nM}}}$$

U_1、U_2、U_3渐近服从正态分布。并且可以确定均匀性假设的临界区间$\overline{R}$，即

$$\overline{R}_{\overline{m}_1}:\left(\frac{1}{2}-U_a\sqrt{\frac{1}{12nM}},\frac{1}{2}+U_a\sqrt{\frac{1}{12nM}}\right) \tag{3-2-16}$$

$$\overline{R}_{\overline{m}_2}:\left(\frac{1}{3}-U_a\sqrt{\frac{4}{45nM}},\frac{1}{3}+U_a\sqrt{\frac{4}{45nM}}\right) \tag{3-2-17}$$

$$\overline{R}_{\overline{s}^2}:\left(\frac{1}{12}-U_a\sqrt{\frac{1}{180nM}},\frac{1}{12}+U_a\sqrt{\frac{1}{180nM}}\right) \tag{3-2-18}$$

当$\overline{m}_1$，$\overline{m}_2$，$\overline{S^2}$大于对应的$\overline{R}$的上界或小于$\overline{R}$的下界，则应否认均匀性假设。U_a可从正态分布的双侧分位数表查得。

（2）均匀性检验　这种检验包括频率检验与累积频率检验。

①频率检验　频率检验亦称拟合优度检验。如果伪随机数是均匀分布的，则可将〔0，1〕分成k个间距相等的子区间，一般可令$k=8$，16，32。此时可作假设H_0：每个伪随机数属于第i个区间的概率为$p_i=\frac{1}{k}$（而$\sum_{i=1}^{k}p_i=1$）。也就是说，频率检验的目的在于检验每个区间观测频数n_i与理论频数$m_i=\frac{n}{k}$之间差别的显著性。为此，可建立x^2检验统计量，即

$$x^2=\frac{k}{n}\sum_{i=1}^{k}\left(n_i-\frac{n}{k}\right)^2 \tag{3-2-19}$$

式中　n——被检验的随机数个数；

k——〔0，1〕上的子区间数；

n_i——落入第i个子区间中的伪随机数个数。该统计量的自由度为$k-1$。

一般情况下，由于仅在一次实验中小概率事件是不容易出现的；如若出现，则认为实际观测频数与理论频数之间相差显著，因而，可以否认假设H_0，亦即认为随机数序列的分布是不均匀的。检验时置信水平可取$a=0.05$或$a=0.01$。若$x^2\geqslant x^2_{0.05}$称为差异显著，反之为差异不显著；若$x^2\geqslant x^2_{0.01}$称为差异极显著，即随机数序列分布是极不均匀的。

②累积频率检验　累积频率检验亦称柯尔莫果洛夫拟合优度检验。如果$F(\gamma)$是伪随机数序列的分布函数，而$S_n(\gamma)$是对该序列作n次独立观测获得的经验分布函数。根据柯尔莫果洛夫——斯米尔诺夫定理，对于任意$\lambda>0$，有等式

$$\lim_{n\to\infty}Q_n(\lambda)=\lim_{n\to\infty}p\left(D_n<\frac{\lambda}{\sqrt{n}}\right)=Q(\lambda) \tag{3-2-20}$$

其中

$$D_n=\sup_{0\leqslant\gamma\leqslant1}|F(\gamma)-S_n(\gamma)|$$

$$Q(\lambda)=\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2\lambda^2}$$

因此，当实验次数n足够大时，就可以认为$D_n<\frac{\lambda}{\sqrt{n}}$的概率$p\left(D_n<\frac{\lambda}{\sqrt{n}}\right)$趋于$Q(\lambda)$。

如果$D_n^{(0)}$是n次实验的$|F(\gamma)-S_n(\gamma)|$中最大的一个，并且$\lambda_0=\sqrt{n}D_n^{(0)}$，当

$$p(\sqrt{n}D_n\geqslant\lambda_0)=1-Q(\lambda_0)=\alpha$$

而α很小时，就出现了小概率事件，由此即可检验差异的显著性。也就是说$\lambda_n\geqslant\lambda_{0.05}$时，则认为伪随机数序列的不均匀性是显著的，否则应当认为序列是均匀的。

（3）独立性检验　独立性检验亦称不相关性检验，是检验伪随机数先后次序之间是否存在相关关系的一种检验方法。独立性检验包括多种具体检验方法，在此仅介绍简单独立性检验及顺序检验。

①简单独立性检验　简单独立性检验又称无重复列联检验。进行这种检验时，首先是将被检验的伪随机数序列划分为ξ与η两部分，并且要求ξ与η中所包含的伪随机数的个数相等，被检验序列中的任一伪随机数都要被取到，而且任一伪随机数只能唯一地属于ξ或η。

设ξ、η的取值分别为ξ_t、$\eta_t(t=1,\ 2,\ \cdots)$，将单位正方形横向分成h列进行ξ取值，而纵向分成k行进行η取值。点$(\xi_t,\ \eta_t)$落入网格$(i,\ j)$内的频数记为n_{ij}。若ξ与η是相互独立的，则ξ_i、η_i同时出现的概率显然为

$$p(\xi_i,\ \eta_i)=p(\xi_i)p(\eta_i) \tag{3-2-21}$$

因而，简单独立性检验就是以（3-2-21）式为假设H_0，比较观测频数n_{ij}与理论频数m_{ij}之间的差异是否显著的一种检验。记

$$\sum_j n_{ij}=n_{i.}$$

$$\sum_i n_{ij}=n_{.j}$$

$$\sum_{i,j} n_{ij}=n$$

检验假设H_0：$p(\xi_i,\ \eta_j)=p(\xi_i)p(\eta_j)$。建立检验统计量$x^2$

$$x^2=\sum_i\sum_j\frac{(n_{ij}-m_{ij})^2}{m_{ij}}$$

即
$$x^2=\sum_i\sum_j\frac{n_{ij}^2}{m_{ij}}-n \tag{3-2-22}$$

因为
$$p(\xi_i)\approx\frac{n_{i.}}{n}$$

$$p(\eta_j)\approx\frac{n_{.j}}{n}$$

所以
$$m_{ij}=np(\xi_i,\ \eta_j)\approx\frac{n_{i.}n_{.j}}{n}$$

又因 $$\sum_j n_{ij} = \sum_j m_{ij} = n_{i.} \qquad (i=1,2,\cdots,h)$$

$$\sum_i n_{ij} = \sum_i m_{ij} = n_{.j} \qquad (j=1,2,\cdots,k)$$

所以自由度为 $$hk-(h+k-1)=(h-1)(k-1) \qquad (3\text{-}2\text{-}23)$$

由（3-2-22）式得到x_0^2值之后，便可与x^2进行比较，若$x_0^2 \geqslant x_{0.05}^2$，则称$\xi$与$\eta$之间显著相关，否则可认为$\xi$与$\eta$之间不相关；若$x_0^2 \geqslant x_{0.01}^2$称$\xi$与$\eta$之间极显著相关。

②顺序检验　顺序检验又称有重复列联检验。如果伪随机数序列是随机的，就不会出现在某一种类型的伪随机数之后总是出现另一种类型伪随机数的现象。

如果将序列中相应的两个随机数组成一个数字，便可组成二元频数表，而所有网格内应当具有近似相等的频数。被检验的n个伪随机数可以组成n对伪随机数，并且$n_{i.}=n_{.i}$，若m_{ij}是n_{ij}的理论频数，则有

$$x^2=\delta_2^2-\delta_1^2 \qquad (3\text{-}2\text{-}24)$$

式中 $$\delta_1^2=\sum_{i=1}^{h}\frac{(n_{i.}-m)^2}{m}, \qquad m=\frac{n}{h}，自由度为h-1；$$

$$\delta_2^2=\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{(n_{ij}-m)^2}{m}, \qquad m=\frac{n}{hk}，自由度为hk-1。$$

上面两式中h、k的意义与前面简单独立性检验时相同。为了计算δ_1^2、δ_2^2，需要将n个伪随机数按大小分为N种类型，即

$$\frac{j-1}{N}\leqslant \gamma_i < \frac{j}{N} \qquad (j=1,2,\cdots,N)$$

$$\frac{k-1}{N}\leqslant \gamma_{i+1} < \frac{k}{N} \qquad (k=1,2,\cdots,N)$$

分类后便可以计算出δ_1^2、δ_2^2以及x^2.若$x_0^2 \geqslant x_{0.05}^2$，则称为显著顺序相关，否则认为不存在顺序相关；若$x_0^2 \geqslant x_{0.01}^2$，则称极显著顺序相关。

（4）组合规律性检验　这种检验是将伪随机数按某种方式进行组合，并且比较各类组合的观测频数n_i与理论频数m_i之间的差别是否显著的一种检验方法。

最常用的是分拆检验，亦称扑克检验，扑克检验一词是由博弈中的花色组合而得名。进行分拆检验时，首先要将伪随机数序列分割为若干个相等的段，一般每个段由k个(如$k=5$)伪随机数组成，分拆时可将k个伪随机数的第一位数字组合起来，并以这种组合上的差别进行分类。例如，k个伪随机数的第一位数字完全相同，则可称为同花；k-1个数字完全相同可称k-1种同花；k个数字互不相同可称k色等等。

当k值较小时，可以分拆出所有可能的组合类型。随着k值增大，各种可能的分拆类型将迅速增加，此时可对分拆简化。Butoher提出了一种比较简单而适用的分拆方法，即每段内的伪随机数为k个时，则可分为1色，2色，…，r色，…，k色总共k种类型。其中的r色就是k个数字中恰有r个不同数字的所有各种组合。而出现r色的概率p_r为（$r\neq 1$）

$$p_r=\frac{a_r(s-1)(s-2)\cdots(s-r+1)}{s^{k-1}}$$

$$p_1=\frac{d_1}{s^{k-1}}$$

由于 $\sum p_r=1$，所以

$$s^k=a_1s+a_2s(s-1)+\cdots+a_ms(s-1)\cdots(s-k+1) \qquad (3\text{-}2\text{-}25)$$

（3-2-25）式中的a_1，a_2，…，a_m由$s=1$，2，…，k依次代入（3-2-25）式而定。

如果按上述分类后，n_i是第i种花色的观测频数，m_i是与n_i对应的理论频数，便可以建立统计量

$$x^2=\sum_{i=1}^{k}\frac{(n_i-m_i)^2}{m_i} \qquad (3\text{-}2\text{-}26)$$

其自由度为$k-1$。

若$x_0^2\geqslant x_{0.05}^2$，则称差异显著，否则认为分拆检验合格；若$x_0^2\geqslant x_{0.01}^2$，则称差异极显著。

（5）无连贯性检验　这种检验是考察在伪随机序列中是否存在某一段的伪随机数普遍偏大，而另一段又普遍偏小；或者某一段的随机数有一贯上升，而另一段又一贯下降等现象。因而，无连贯性检验就是对伪随机数出现先后顺序的随机性检验。

最常用的是两类连检验，这种检验是以$q=1-p$为基准对伪随机数进行分类，若$\gamma_i>q$，则归入A类，若$\gamma_i\leqslant q$，则归入B类。这样就得到性质不同的两类A与B，所谓连就是指连续出现的同类伪随机数构成的一个段，段中包含的伪随机数个数称为连的长度，简称连长，以k表示。对于由n个伪随机数组成的序列，m、n分别表示A、B两类包含的伪随机数个数。l_{1i}表示长度为i的A类连数；l_{2i}表示长度为i的B类连数；$l_i=l_{1i}+l_{2i}$表示长度为i的总连数。则有

$L_{1k}=\sum_{i=k}^{m}l_{1i}$表示连长不小于$k$的$A$类连数；

$L_{2k}=\sum_{i=k}^{m}l_{2i}$表示连长不小于$k$的$B$类连数；

$L_k\;=L_{1k}+L_{2k}$表示连长不小于k的总连数；

$L_1\;=\sum_{i=1}^{m}l_{1i}$表示$A$类的连数；

$L_2\;=\sum_{i=1}^{m}l_{2i}$表示$B$类的连数；

$L\;=L_1+L_2$表示总连数。

对于10进制的伪随机数，可把〔0，1〕划分为10个相等的子区间，此时可把伪随机数按它的第一位数字分成10类。分类后，伪随机数序列就可划分为不同长度的A、B两类连。

如果按$q=1-p$对伪随机数分类，其中的$0<q<1$，则A、B两类随机数出现的概率分别为p和q，$p+q=1$。若由n个伪随机数组成的序列是局部随机的，则出现A、B两类个数的数学期望分别为n_p、n_q；而在A、B类之后，出现长度为k的B、A类连的概率分别为q^kp、p^kq，$k=1$，2，…；理论频率分别为Nq^kp、Np^kq。至此，便可以进行x^2检验，以确定伪随机数序列的随机性。

前面简要地介绍了均匀分布伪随机数的产生与检验。应用蒙特卡罗法对石油资源量进行模拟计算时，除用均匀分布伪随机数外，有时也用到正态分布伪随机数或其他种分布的伪随机数。而这些各种不同分布的伪随机数，都可以用均匀分布伪随机数经过数学变换得到。

例如，为了得到正态分布的伪随机数序列，可取均匀分布伪随机数序列中的γ_i及γ_{i+1}，经过如下变换便可以得到正态分布伪随机数序列中的u_i及u_{i+1}。

$$\begin{cases} u_i=\sqrt{-2\ln\gamma_i}\ \cos2\pi\gamma_{i+1} \\ u_{i+1}=\sqrt{-2\ln\gamma_i}\ \sin2\pi\gamma_{i+1} \end{cases} \tag{3-2-27}$$

上式中，u_i及u_{i+1}分别为正态分布伪随机数序列中的第i个及第$i+1$个伪随机数。

二、构造随机变量的分布函数

石油资源量（或储量）计算公式中的参数，一种是常数（经验系数或地质常数），一种是随机变量。对于随机变量首先要构造出它的分布函数。

随机变量X的值小于实数x的概率$p(X<x)$，是x的函数，记作$F(x)=p(X<x)$，函数$F(x)$叫作随机变量X的分布函数。因为在估算石油资源量时，总是希望知道在100%概率下的石油资源量有多少，所以，在石油资源评价中是把随机变量的分布函数定义为$AF(x)=1-F(x)=P(X>x)$。$AF(x)$与$F(x)$的曲线形态反向对称，见图3-2-1。

石油勘探阶段，特别是早期勘探阶段，所能收集到的地质参数的数量一般都比较少。因而，构造随机变量的分布函数$AF(x)$时，要根据地质参数的数量多少，分别采用不同的方法进行处理。

经常遇到的情况有如下三种：①原始数据的数量大于30个的大子样类型；②原始数据的数量小于30个的小子样类型；③原始数据的数量极少，而且又不知道随机变量的分布概型。

需要指出的是，根据原始数据经过计算构造出的分布函数，都是以有限个的离散值存入计算机中，而不可能是连续的分布函数。

1. 用频率统计法构造经验分布函数

当随机变量的原始数据为大子样时，可使用频率统计法构造分布函数。这种方法就是通

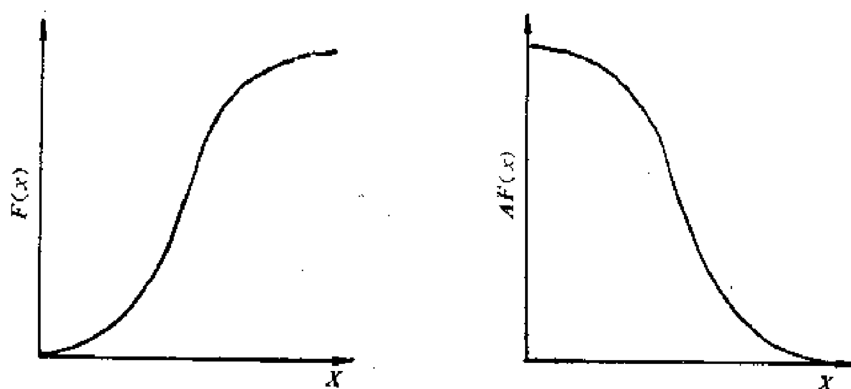


图3-2-1 $F(x)$与$AF(x)$的曲线形态

常所说的由观测数据构造直方图的方法。

用这种方法求得的分布函数，由于直接来自实际观测资料，所以一般称为经验分布函数。只要观测数据的代表性较好，多数情况下，经验分布函数的代表性也较好，而且不受各种理论分布函数概型所约束，使用方便，效果也很好。

用频率统计法求经验分布函数时，统计频率的区间个数最好是奇数，这样可保证随机变量密度分布的峰值出现。频率统计区间的个数可按如下原则确定，即平均落入每个区间的原始数据数量不少于3～5个。

在程序设计上，应该使计算机在计算时，能根据原始数据的数量，按上述原则自动地确定出最佳的区间个数M，并按（3-2-28）式求出$M+1$个区间间隔值X_i。

$$X_i = X_{min} + (XL/M)(i-1)$$
$$= X_{min} + DX(i-1) \qquad (i=1, 2, \cdots, M+1) \qquad (3\text{-}2\text{-}28)$$

式中　X_i——第i个分隔值；

X_{min}——原始数据中的最小值；

X_{max}——原始数据中的最大值；

XL——极差，$XL = X_{max} - X_{min}$；

DX——区间增量。

例如，某地质凹陷中下第三系的某一储油层，经过钻井勘探已见到油气显示。为估计其含油前景，需要研究储集层的厚度（H）参数。根据地震资料加上少数钻井资料校正，总共得到46个厚度数据。这46个数据就是推测母体分布的一个大子样，见表3-2-2。

表3-2-2　储集层厚度数据表

序　号	厚　度（m）	序　号	厚　度（m）	序　号	厚　度（m）	序号	厚　度（m）
1	20.0	13	38.5	25	54.0	37	79.5
2	46.5	14	63.0	26	31.5	38	83.0
3	39.5	15	47.0	27	68.0	39	58.0
4	46.0	16	65.0	28	36.5	40	61.5
5	29.0	17	35.0	29	50.0	41	53.5
6	51.0	18	42.5	30	49.0	42	60.5
7	41.5	19	70.5	31	64.5	43	4[illegible].5
8	52.0	20	43.5	32	75.5	44	51.5
9	42.0	21	53.0	33	37.5	45	52.5
10	59.0	22	30.0	34	56.0	46	28.5
11	47.5	23	40.5	35	57.0		
12	62.5	24	70.5	36	25.5		

表3-2-2中的最大值$X_{max}=83.0$m，最小值$X_{min}=20.0$m，极差$XL=83.0-20.0=63.0$m。按平均落入每个频率统计区间的样品数为5个计算，46个原始数据应分为9个区间，区间增量$DX = XL/M = 63.0/9 = 7.0$m。按（3-2-28）式可求出9个区间的10个间隔值如下：

20.0　　27.0　　34.0　　41.0　　48.0

55.0　　62.0　　69.0　　76.0　　83.0

为求储集层厚度的分布函数，可先作出储集层厚度的频率直方图，见图3-2-2。

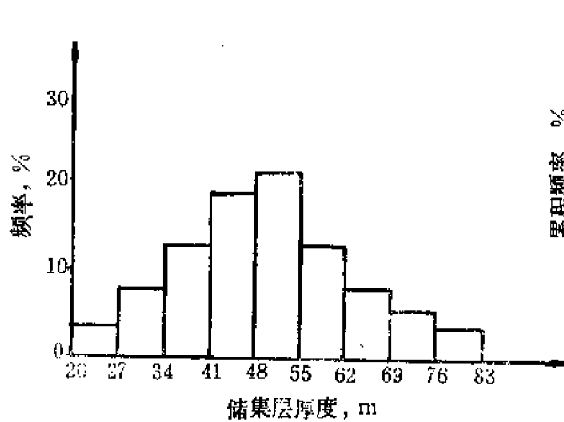


图3-2-2　储集层厚度的频率直方图

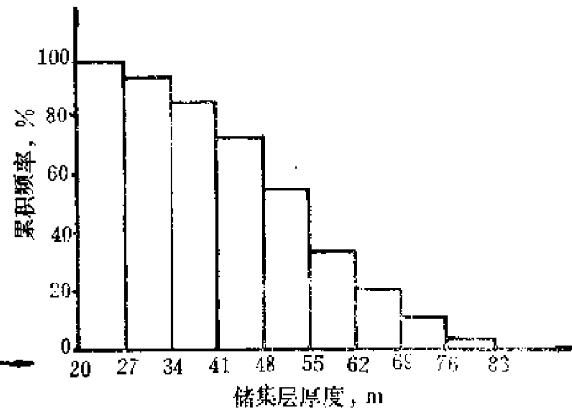


图3-2-3　储集层厚度的累积频率直方图

如果从第9个区间开始向左逐个区间累加频率值，可以得到累积频率直方图，见图3-2-3。

把累积频率直方图上的每个区间中点连线，便可以得到储集层厚度的累积频率折线图，见图3-2-4。

这个仅由9个点组成的折线图形，可以看作是储集层厚度这一随机变量 H 的分布函数 $AF(h)$ 的近似图形。

表3-2-3中给出了储集层厚度9个区间中原始数据出现的频数、频率、累积频率。其中，累积频率的9个值应存入计算机，以备尔后抽样计算时使用。

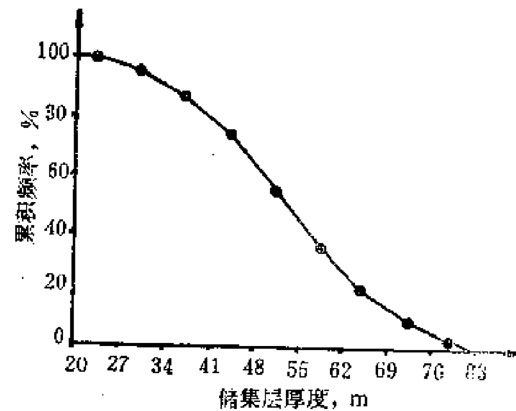


图3-2-4　储集层厚度累积频率折线图

表3-2-3　储集层厚度的累积频率数据表

序号	区间间隔值 (m)	频率	频率 (%)	累积频率 (%)
1	20.0～27.0	2	0.043	1.000
2	27.0～34.0	4	0.087	0.957
3	34.0～41.0	6	0.130	0.870
4	41.0～48.0	9	0.195	0.739
5	48.0～55.0	10	0.217	0.543
6	55.0～62.0	6	0.130	0.326
7	62.0～69.0	4	0.087	0.196
8	69.0～76.0	3	0.065	0.109
9	76.0～83.0	2	0.043	0.043

2. 用样品等频率法构造经验分布函数

当随机变量的原始数据量为10～30个，而且又不知道随机变量的分布概型时，如果采用频率统计法，则因统计区间个数过少，而使构造的分布函数过于粗糙；如果采用下面的以假定分布概型代替，又感到没有充分发挥每个原始数据的作用。在这种情况下，可用样品等频率法构造经验分布函数。

样品等频率法的出发点是认为每个样品（随机变量的数据）出现的概率是相等的，如果子样的容量为N，则每个样品的出现概率为$1/N$。若把N个观测值按大小排列，可以得到递增序列X_1，$X_2,\cdots,X_N$，这N个观测值对应的频率值$AF(x_i)$可由下面的公式（3-2-29）求出：

$$AF(x_i)=\frac{N-i}{N-1} \qquad (i=1,2,\cdots,N) \tag{3-2-29}$$

式中 $AF(x_i)$——递增序列中第i个观测值对应的累积频率值；

N——子样容量。

例如，为了计算某一油田的地质储量，我们根据表3-2-4列出的仅有的11个储集层孔隙度（ϕ）数据，构造经验分布函数。

表3-2-4 储集层孔隙度数据表

序 号	孔隙度（%）	序 号	孔隙度（%）	序 号	孔隙度（%）	序 号	孔隙度（%）
1	16.7	4	14.0	7	12.6	10	14.6
2	13.4	5	8.0	8	15.4	11	10.4
3	18.0	6	11.7	9	20.0		

表3-2-4中的孔隙度数据由小到大重新排序后如下：

8.0　10.4　11.7　12.6　13.4　14.0

14.6　15.4　16.7　18.0　20.0

因为子样容量为$N=11$，所以可将频率坐标轴分为10个等间隔区间，区间的频率增量为10%。在孔隙度坐标轴上找到11个原始数据的位置，这样便可以得到11个孔隙度ϕ与累积频率$AF(\varphi)$对应的坐标点$(\phi_i\ AF(\varphi_i))$。例如，排序后的第1个样品的坐标点为（8.0，100%），第6个样品的坐标点为（14.0，50%），第11个样品的坐标点为（20.0，0%）。把这11个点用直线连接起来便可以得到累积频率折线图，见图3-2-5。

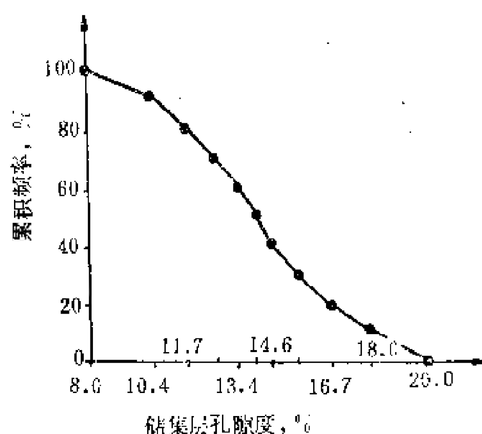


图3-2-5 储集层孔隙度累积频率折线图

图3-2-5中的累积频率折线，可以看作是储集层孔隙度ϕ这一随机变量的分布函数$AF(\varphi)$的近似图形。表3-2-5中列出了储集层孔隙度11个原始数据所对应的累积频率值。

3. 用理论分布概型公式构造分布函数

当随机变量的原始数据量为小子样，但我们知

表3-2-5　储集层孔隙度的累积频率数据表

序号	孔隙度（%）	累积频率（%）	序号	孔隙度（%）	累积频率（%）	序号	孔隙度（%）	累积频率（%）
1	8.0	100	5	13.4	60	9	16.7	20
2	10.4	90	6	14.0	50	10	18.0	10
3	11.7	80	7	14.6	40	11	20.0	0
4	12.6	70	8	15.4	30			

道这一分布函数符合或接近某种分布的理论概型时，则可用这种理论分布概型公式构造随机变量的分布函数。

根据大量的实际资料统计结果，计算石油资源量（或储量）的许多参数的分布函数都符合或接近正态分布或对数正态分布。例如，一个探区的地层厚度、储集层孔隙度等大都符合正态分布；油田储量、地质圈闭面积的对数值等大都符合正态分布。

例如，用容积法计算石油地质储量时需要含油饱和度这一地质参数，由于含油饱和度这一参数难以求准，目前多用（1－含水饱和度）代替含油饱和度。经过化验室分析结果，得到5个储集层含水饱和度数据：6.0%，7.0%，8.0%，4.0%，5.0%。根据与含油气地质条件相似的邻区对比，知道含水饱和度的分布函数接近正态分布。

为了求得含水饱和度的分布函数，首先需要求出这5个数据的平均值$\overline{X}$及标准差s。

$$\overline{X}=\frac{1}{5}\sum_{i=1}^{5}X_i=6.0\%$$

$$s=\sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(X_i-6.0\%)}=1.581\%$$

知道了平均值$\overline{X}$及标准差s之后，便可按正态分布的理论公式求出其密度函数$f(x)$，对密度函数进行积分便可以得到分布函数$AF(x)$。

$$f(x)=\frac{1}{\sqrt{2\pi}\,s}\,e^{-\frac{(x-\overline{X})^2}{2s^2}} \qquad (3\text{-}2\text{-}30)$$

$$AF(x)=1-\frac{1}{\sqrt{2\pi}\,s}\int_{-\infty}^{x}e^{-\frac{(x-\overline{X})^2}{2s^2}}dx \qquad (3\text{-}2\text{-}31)$$

（3-2-31）式积分后，随机变量X变化范围的左、右界线u_1、u_2可按3倍标准差的原则取值，即

$$u_1=\overline{X}-3s$$
$$u_2=\overline{X}+3s$$

但是，用这种方法构造的分布函数，其随机变量的变化范围将比实际观测值范围增大。如上例中含水饱和度的实际观测值变化范围是4.0～8.0%，但是按3倍标准差构造的正态分布函数，其含水饱和度的变化范围却增大为1.26～10.74%。这种情况称为随机变量的区间增值，有时增值会使随机变量的范围过大，甚至会出现不合理的负值。所以，随机变量X的左右界限，要由对探区地质情况比较了解的地质家根据参数的可能变化范围确定，也可以参

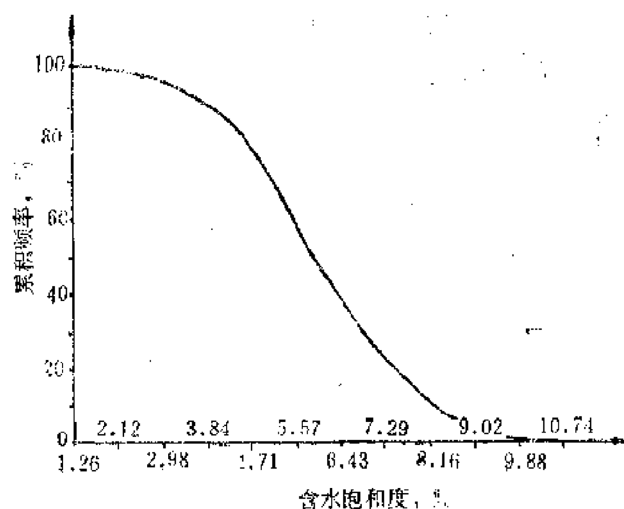


图3-2-6　含水饱和度的正态分布函数

考邻区的参数变化范围。

按正态分布概型构造的含水饱和度的分布函数见图3-2-6。表3-2-6中给出了含水饱和度正态分布函数的45个数值点。

表3-2-6　含水饱和度正态分布函数的数据表

序号	含水饱和度（%）	累积频率（%）	序号	含水饱和度（%）	累积频率（%）	序号	含水饱和度（%）	累积频率（%）	序号	含水饱和度（%）	累积频率（%）
1	1.2566	1.0000	13	3.8439	0.9136	25	6.4512	0.3023	37	9.0185	0.0281
2	1.4722	0.9979	14	4.0595	0.8900	26	6.6468	0.3410	38	9.2341	0.0204
3	1.6878	0.9963	15	4.2751	0.8622	27	6.8624	0.2925	39	9.4498	0.0145
4	1.9034	0.9952	16	4.4907	0.8399	28	7.0780	0.2475	40	9.6054	0.0102
5	2.1190	0.9929	17	4.7063	0.7932	29	7.2937	0.2064	41	9.8810	0.0070
6	2.3346	0.9893	18	4.9220	0.7521	30	7.5093	0.1697	42	10.09[illegible]	0.0048
7	2.5502	0.9854	19	5.1376	0.7070	31	7.7249	0.1375	43	10.3122	0.0032
8	2.7659	0.9796	20	5.3532	0.6585	32	7.9405	0.1097	44	10.5278	0.0021
9	2.9815	0.9718	21	5.5688	0.6072	33	8.1561	0.0862	45	10.7434	0.0000
10	3.1971	0.9618	22	5.7844	0.5540	34	8.3717	0.0667			
11	3.4127	0.9490	23	6.0000	0.4997	35	8.5873	0.0508			
12	3.6283	0.9331	24	6.2156	0.4[illegible]55	36	8.8029	0.0381			

4. 用三角分布构造分布函数

当随机变量的原始数据只有3个：一个是最大值X_3，一个是最小值X_1，一个是介于X_3与X_1之间的最可能值X_2，而且又不知道随机变量的分布概型时，可用三角分布代替随机变量的分布函数：

$$AF(x)=\begin{cases}1-\dfrac{(x-X_1)^2}{(X_3-X_1)(X_2-X_1)} & (x\leqslant X_2)\\[2ex] \dfrac{(X_3-x)^2}{(X_3-X_1)(X_3-X_2)} & (x>X_2)\end{cases} \tag{3-2-32}$$

用三角分布构造的分布函数，当最可能值X_2位于X_3与X_1的中点时，曲线形态与正态分布相近；而当X_2不在中点时，曲线形态则为偏态分布。

近年来，三角分布在海洋石油勘探的油气资源评价中已得到相当广泛的应用，显然是因为海洋钻井费用昂贵，资料较少所造成的。前已述及，由于计算石油资源量的许多地质参数符合或接近正态分布，而三角分布恰与正态分布相近，因而，在原始数据极少的条件下，用三角分布构造随机变量的分布函数是一种可行的办法。

例如，在某海洋大陆架发现一个含油面积较大的油田，但只有3个岩心含水饱和度的实验分析数据，最大值为4.0%，最小值为2.0%，最可能值为3.0%。

按（3-2-32）式的三角分布公式计算可以得到形态光滑的分布函数曲线，见图3-2-7。表3-2-7给出了含水饱和度三角分布函数的45个数值点。

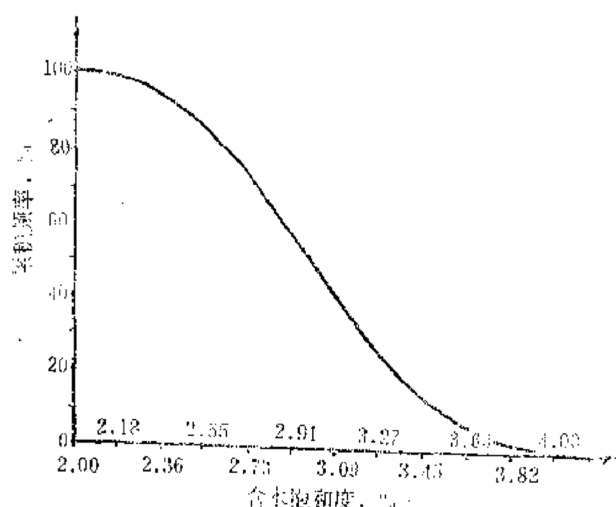


图3-2-7　含水饱和度的三角分布函数

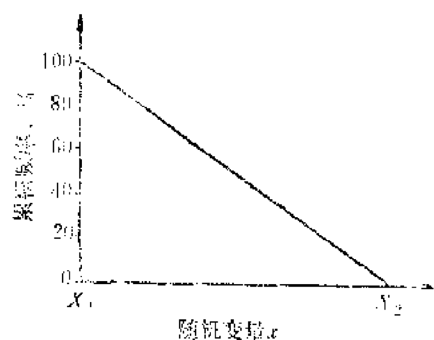


图3-2-8　均匀分布函数

5、用均匀分布构造分布函数

当随机变量的原始数据只有2个，一个是最大值X_2，一个是最小值X_1，而且又不知道随机变量的分布概型，在这种情况下可用最简单的均匀分布代替随机变量的分布函数。但是，这种作法应当尽量回避，因为地质上的随机变量服从均匀分布的较少。均匀分布函数的公式为

$$AF(x)=1-\frac{x-X_1}{X_2-X_1}=\frac{X_2-x}{X_2-X_1} \tag{3-2-33}$$

均匀分布函数$AF(x)$的图形见图3-2-8。

表3-2-7　含水饱和度三角分布函数的数据表

序号	含水饱和度（%）	累积频率（%）	序号	含水饱和度（%）	累积频率（%）	序号	含水饱和度（%）	累积频率（%）	序号	含水饱和度（%）	累积频率（%）
1	2.0000	1.0000	13	2.5455	0.8512	25	3.0909	0.4132	37	3.6364	0.0601
2	2.0455	0.9990	14	2.5909	0.8254	26	3.1364	0.3729	38	3.6818	0.0506
3	2.0909	0.9959	15	2.6364	0.7975	27	3.1818	0.3347	39	3.7273	0.0372
4	2.1364	0.9907	16	2.6818	0.7676	28	3.2273	0.2986	40	3.7727	0.0258
5	2.1818	0.9835	17	2.7273	0.7355	29	3.2727	0.2645	41	3.8182	0.0165
6	2.2273	0.9742	18	2.7727	0.7014	30	3.3182	0.2324	42	3.8636	0.0093
7	2.2727	0.9628	19	2.8182	0.6653	31	3.3636	0.2025	43	3.9091	0.0041
8	2.3182	0.9494	20	2.8636	0.6271	32	3.4091	0.1746	44	3.9545	0.0010
9	2.3636	0.9339	21	2.9091	0.5868	33	3.4545	0.1488	45	4.0000	0.0000
10	2.4091	0.9163	22	2.9545	0.5444	34	3.5000	0.1250			
11	2.4545	0.8967	23	3.0000	0.5000	35	3.5455	0.1033			
12	2.5000	0.8750	24	3.0455	0.4556	36	3.5909	0.0837			

三、计算局部含油地质单元的石油资源量

局部含油地质单元是计算石油资源量或石油储量的基本地质体，采用不同计算方法时，局部含油地质单元的含义可能不一致。例如，采用容积法计算石油储量时，局部含油地质单元可以是一个油藏或一个油层；采用单储系数法计算石油资源量时，局部含油地质单元可以是局部构造、断鼻、断块；采用氯仿沥青法计算石油资源量时，局部含油地质单元可以是一个生油凹陷，等等。

如果某个含油区中总共有m个局部含油地质单元，其中第j个局部含油地质单元石油资源量的计算公式可用（3-2-1）式表示。若在计算公式中有n个地质参数，其中有t个随机变量，有（$n-t$）个常数或经验系数，则第j个局部含油地质单元的石油资源量Q_j为

$$Q_j=\prod_{i=1}^{n}X_{ij}=\prod_{i=1}^{t}X_{ij}\cdot\prod_{i=t+1}^{n}X_{ij}$$

$$=K\prod_{i=1}^{t}X_{ij}\qquad(i=1,2,\cdots,t)(j=1,2,\cdots,m)\tag{3-2-34}$$

上式中的$\prod_{i=t+1}^{n}X_{ij}$为$(n-t)$个常数或经验系数的连乘积，仍为常数，令其为K。

局部含油地质单元的石油资源量计算公式可以形象化地表现为如下形式，见图3-2-9。

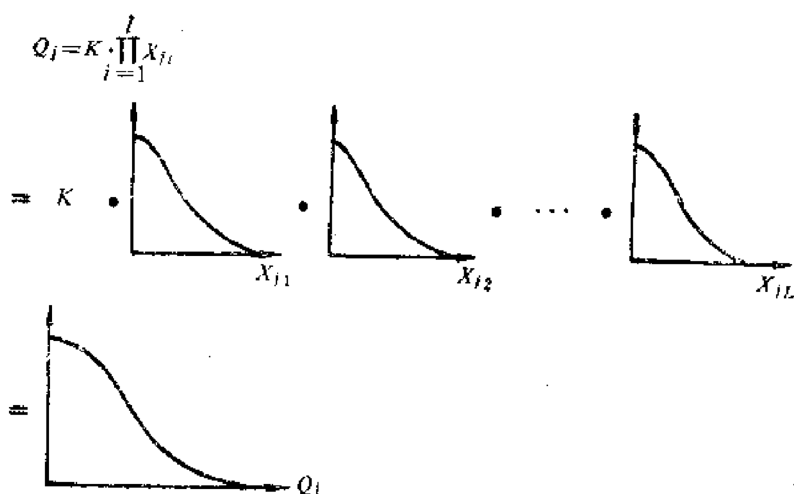


图3-2-9　计算局部含油地质单元的石油资源量示意图

从图3-2-9可以看出，计算石油资源量Q_j的技术关键是如何实现随机变量X_{ji}之间的乘法运算。这里介绍的具体实现方法是以［0，1］区间上均匀分布的随机数作为抽样序列，用直接抽样法完成模拟计算。

在模拟计算之前，首先需要求出第j个局部含油地质单元石油资源量Q_j出现的最大可能范围，即确定资源量的最大可能值Q_{max}及最小可能值Q_{min}。这里说成为可能值，其原因是抽样计算后Q_j的变化范围有可能要缩小，即计算后的Q_{jmax}<计算前的Q_{jmax}，计算后的Q_{jmin}>计算前的Q_{jmin}。这个情况以后再作说明。

$$Q_{jmax} = K\prod_{i=1}^{l} X_{jimax} \qquad (j=1,2,\cdots,m) \qquad (3\text{-}2\text{-}35)$$

$$Q_{jmin} = K\prod_{i=1}^{l} X_{jimin} \qquad (j=1,2,\cdots,m) \qquad (3\text{-}2\text{-}36)$$

（3-2-35）及（3-2-36）式中的X_{jimax}及X_{jimin}分别是石油资源量计算公式中的第i个随机变量的极大值及极小值。

$$QL_j = Q_{jmax} - Q_{jmin} \qquad (j=1,2,\cdots,m) \qquad (3\text{-}2\text{-}37)$$

（3-2-37）式中的QL_j为资源量Q_j出现的最大可能范围。

在计算时为了统计抽样值的累积频率，需要把QL_j分为w个区间，一般情况下可令w=100，那么，$w+1$个区间间隔值为

$$Q_{jh} = Q_{jmin} + (QL_j/w)(h-1) \qquad (j=1,2,\cdots,m)(h=1,2,\cdots,w+1) \quad (3\text{-}2\text{-}38)$$

抽样模拟计算的具体步骤如下：

首先以［0，1］区间上均匀分布的随机数序列中的第s个随机数γ_s，作为第i个随机变量

X_{ji}的分布函数$AF(x_{ji})$的概率入口值，沿X_{ji}轴方向作平行线，与分布函数相交后，再沿$AF(x_{ji})$轴方向作平行线，直到与X_{ji}轴相交并得到交点X_{jiw}，X_{jiw}称为出口值。而实际计算时，是用插值法（线性或非线性插值）求出随机变量X_{ji}的出口值X_{jiw}。如果分布函数$AF(x_{ji})$有u个离散值，若按线性插值方法计算时，则有

$$X_{jiw}=\frac{(X_{jip}-X_{jip-1})[\gamma_s-AF(x_{jip-1})]}{[AF(x_{jip})-AF(x_{jip-1})]}+X_{jip-1}$$

$(j=1,2,\cdots,m)(i=1,2,\cdots,t)(p=1,2,\cdots,u)(w=1,2,\cdots,g)$ （3-2-39）

上式中的γ_s是随机数序列中序号为s的随机数。

继续以随机数序列中的第$s+1$个随机数γ_{s+1}，作为第$i+1$个随机变量的分布函数$AF(x_{ji+1})$的概率入口值，并用（3-2-39）式求出第$i+1$个随机变量X_{ji+1}的出口值X_{ji+1w}。这一抽样过程可见图3-2-10。

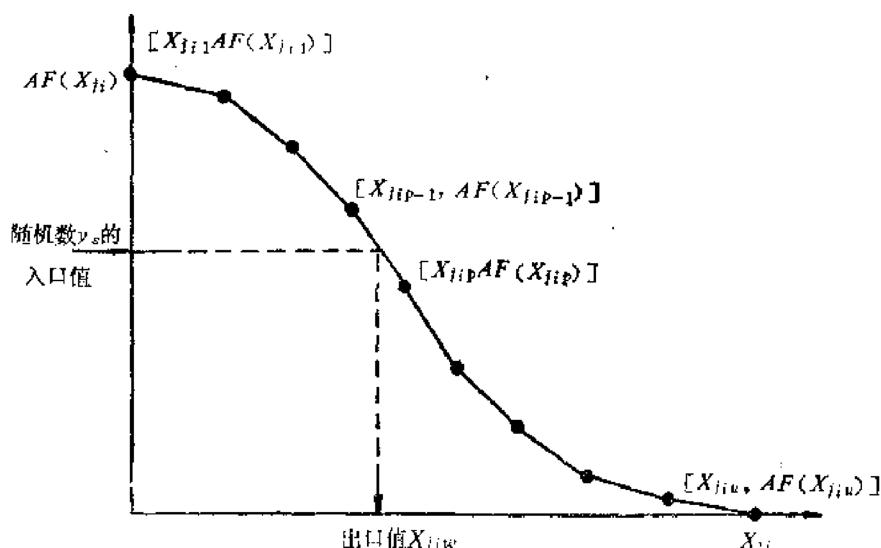


图3-2-10 抽样过程示意图

如此，逐一求出石油资源量计算公式中全部t个随机变量的出口值。

这t个出口值相乘，再乘以资源量计算公式中的常数与经验系数的乘积值K，则得到第j个局部含油地质单元石油资源量的一个随机估计值Q_{jw}，即

$$Q_{jw}=K\prod_{i=1}^{t}X_{jiw}\qquad(j=1,2,\cdots,m)(w=1,2,\cdots,g)$$

重复上述步骤，如果进行了g次，则可得到g个第j个局部含油地质单元石油资源量的随机估计值。在这一模拟过程中，总共使用了gt个随机数。

最后以频率统计法对g个随机估计值进行整理，就可以求出含油区中第j个局部含油地质单元的石油资源量Q_j的分布函数$AF(q_j)$。

抽样次数g值的大小，一般没有明确的标准。理论上要求g值越大越好，在实际计算中可以逐步增加g值，直到再增加g值时，分布函数$AF(q_j)$的曲线形态已不再变化时为止，这时的抽样次数g，即可认为是实际使用时的最佳抽样次数。

当频率统计区间数$k=100$时，按作者多次实践经验，一般情况下抽样次数g可选为500～2000次，最多也不必超过10000次。

四、计算含油区的石油资源总量

这里所说的含油区可以是一个地质凹陷、一个地质坳陷、一个沉积盆地、以至超越盆地的一个范围较大的含油气地区。因此，含油区的石油资源总量可能需要多级累加才能求出。例如，如含油区是一个沉积盆地，局部含油地质单元是地质圈闭的话，则要根据地质圈闭的石油资源量先求出地质凹陷的石油资源量，再由地质凹陷的石油资源量求出地质坳陷的石油资源量，最后再由地质坳陷的石油资源量求出全盆地的石油资源总量，即

$$Q=\sum_{j_1=1}^{m_1}\sum_{j_2=1}^{m_2}\cdots\sum_{j_k=1}^{m_k}Q_{j_1j_2\cdots j_k} \qquad (3\text{-}2\text{-}40)$$

式中　Q——含油区的石油资源总量；

$Q_{j_1j_2\cdots j_k}$——第j_1级分区中的、第j_2分区中的…第j_k个分区中的局部含油地质单元的石油资源量。

为了简化问题，这里仅以局部含油地质单元一次求和作为含油区的石油资源总量。因而，可按前面给出的（3-2-2）式进行计算，而这一计算过程可以形象化地表示为如下形式，见图3-2-11。

从图3-2-11可以看出，计算含油区石油资源总量的技术关键是如何实现局部含油地质单元的石油资源量分布函数之间的加法运算。

在模拟计算之前，也要先求出含油区石油资源总量Q出现的最大可能范围，即确定资源总量Q的最大可能值Q_{max}及最小可能值Q_{min}以及范围值QL

$$Q_{max}=\sum_{j=1}^{m}Q_{jmax} \qquad (3\text{-}2\text{-}41)$$

$$Q_{min}=\sum_{j=1}^{m}Q_{jmin} \qquad (3\text{-}2\text{-}42)$$

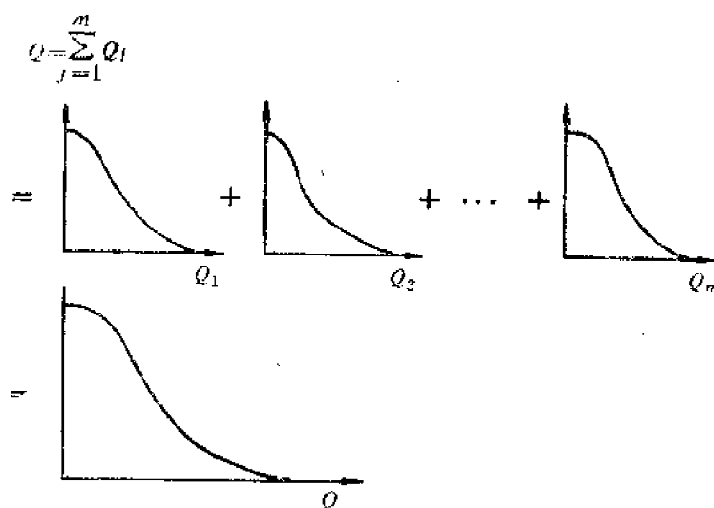


图3-2-11　计算含油区石油资源总量的示意图

$$QL = Q_{max} - Q_{min} \quad (3\text{-}2\text{-}43)$$

再把QL分为若干个区间，例如分为100个子区间，以〔0，1〕区间上均匀分布随机数作为每个局部含油地质单元石油资源量Q_j的分布函数$AF(q_j)$的概率入口值，用插值法求出石油资源量Q_j的出口值。如果含油区中有m个含油地质单元，则累加m个出口值，得到含油区石油资源总量的一个随机估计值。在这一抽样计算过程中，如果抽样g次，则总共使用了gm个随机数。最后以频率统计法就可以求出含油区石油资源总量Q的分布函数$AF(q)$。

五、石油资源总量分布函数的正态化及内插整理

由若干个局部含油地质单元的石油资源量分布函数经过多次加法运算，求得的含油区石油资源总量分布函数$AF(q)$的形态趋向正态分布。产生这种现象的原因是由n个抽样值累加和的均匀化所致，它服从中心极限定理。

根据李雅普诺夫（Liapunov）定理有：如果随机变量Q_1，Q_2，…，Q_n，…相互独立，它们具有有限的数学期望$E(Q_j)$和方差$D(Q_j)$，即

$$E(Q_j) = a_j \quad (j = 1, 2, \cdots, n, \cdots)$$
$$D(Q_j) = \sigma_j^2 \quad (\sigma_j^2 \neq 0) \quad (j = 1, 2, \cdots, n, \cdots)$$

记：
$$B_n^2 = \sum_{j=1}^{n} \sigma_j^2$$

若存在正整数δ，当$n \to \infty$时有

$$\frac{1}{B_n^{2+\delta}} \sum_{j=1}^{n} E|Q_j - a_j|^{2+\delta} \to 0$$

则随机变量
$$Z_n = \left(\sum_{j=1}^{n} Q_j - \sum_{j=1}^{n} a_j\right) / B_n$$

的分布函数$F_n(q)$对于任意q，满足：

$$\lim_{n \to \infty} F_n(q) = \lim_{n \to \infty} p(Z_n \leqslant q) = \int_{-\infty}^{q} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

亦即，当n充分大时，随机变量Z_n将服从正态分布N（0，1）。

由此得出，当n充分大时

$$\sum_{j=1}^{n} Q_j = B_n Z_n + \sum_{j=1}^{n} a_j$$

将服从正态分布$N\left(\sum_{j=1}^{n} a_j,\ \sum_{j=1}^{n} \sigma_j^2\right)$。

这就是说，无论各个局部含油地质单元石油资源量Q_j的分布函数$AF(q_j)$具有怎样的分布，只要满足上述定理的条件，则当累加次数n充分大时，含油区石油资源总量$\sum_{j=1}^{n} Q_j$的分布函数$AF(q)$将会近似地服从正态分布。

此外，还要指出一个重要现象：经过多次累加后的含油区石油资源总量的分布函数，其

资源量的实际变化范围要小于Q_{max}至Q_{min}的范围。为了叙述上的方便，现以$Q=Q_1+Q_2$为例加以说明，见图3-2-12。

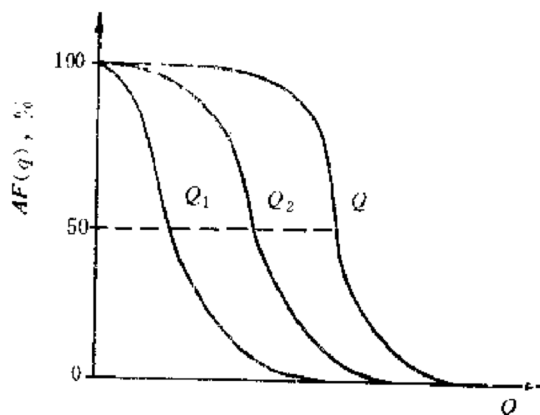


图3-2-12　石油资源总量区间范围收缩现象

假如Q_1及Q_2的分布函数$AF(q_1)$及$AF(q_2)$都属于正态分布，经过抽样值的加法运算后得到Q的分布函数$AF(q)$。此时，只在概率为50%处$Q=Q_1+Q_2$，当概率>50%时，$Q>Q_1+Q_2$，而概率<50%时，$Q<Q_1+Q_2$。

这种现象的出现与否以及发生的程度与两个因素有关，其一是与被累加的分布函数$AF(q_i)$的概型有关；其二是与累加的次数有关。如果被累加的分布函数均为正态分布，则区间两侧的收缩程度相等。而累加的次数越多，则区间收缩得也越明显。所以，在分布区间的石油资源量大值一侧，含油区的石油资源总量Q小于m个局部含油地质单元的石油资源量的累加和，即

$$Q<\sum_{i=1}^{m}Q_i$$

而在区间的石油资源量小值一侧，含油区的石油资源总量Q大于m个局部含油地质单元的石油资源量的累加和，即

$$Q>\sum_{i=1}^{m}Q_i$$

因此，经过抽样模拟计算后得到的含油区石油资源总量，其分布函数$AF(q)$的极大值一端，往往会出现很多个概率为100%的数值点；而在极小值一端，会出现很多个概率为0%的数值点。因为分布函数$AF(q)$的两端只应保留概率为100%及0%的点各一个，为去掉多余的概率为100%及0%的数值点，可通过区间〔Q_{min}，Q_{max}〕内部插值计算，求出有效区间〔Q'_{min}，Q'_{max}〕范围内对应的分布函数$AF'(q)$。

但是，这种内插整理计算，需要在全部累加计算完成后一次进行，以防止多次内插计算时产生误差的累积传播。

六、风险分析

由于石油勘探的未来成效具有不确定性，因而特别需要对估算的石油资源量进行风险分析。所谓风险就是失败的机会。石油勘探中的风险是多种多样的。如勘探地区是否具备形成油气藏地质条件的地质风险；在已具备形成油气藏地质条件的含油区内，经过勘探能否找到一定规模油气藏的勘探风险；勘探后已经发现的油气藏是否具备开采价值的经济风险；石油勘探过程中人与设备是否安全的环境风险；对于勘探地区，特别是海域大陆架地区是否有国际争议的政治风险等等。

显而易见，上述种种风险都会对石油勘探起着决定性影响。因而风险分析是石油资源评价工作的不可缺少的重要环节。

石油地质勘探的专业人员必须认真作好地质风险分析。其他方面的风险分析可由对口的专业人员去完成。在实际工作中，地质风险分析可以在不同的层次进行，例如，单一地质圈闭的风险分析，一组地质圈闭（国外对地质条件相似的一组地质圈闭称作一个勘探层）的风险分析，一个油气聚集带的风险分析，以及整个含油气盆地的风险分析等等。但是，需要明确指出的是，在石油资源量计算过程中，地质风险分析只需要进行一次。

一般情况下，地质风险分析大都从地质圈闭算起，其计算公式为

$$R=1-\prod_{i=1}^{n}(1-r_i) \qquad (3\text{-}2\text{-}44)$$

式中 R——风险值；

r_i——第i个地质因素的风险值。

例如，地质勘探人员用如下的容积法公式估算一个地质圈闭的石油储量

$$Q=SH\Phi DW \qquad (3\text{-}2\text{-}45)$$

式中 S——含油面积；

H——储集层厚度；

Φ——储集层孔隙度；

D——石油充满系数；

W——采收率。

风险分析时，要由熟悉探区情况的地质人员对上述5个地质参数逐个地进行分析论证。就一般的地质概念而论，上述5个地质参数中，含油面积S的风险常常决定于地质调查或地震勘探资料的可靠性；储集层厚度H的风险受岩性岩相变化的影响；储集层孔隙度Φ的风险决定于储集层孔隙是否有次生改造或后期充填的影响；石油充满系数D的风险可能受生油岩的成熟程度及油气运移通道的制约；而采收率W的风险则与原油性质及驱动类型有关。

经过认真分析论证后，要对5个地质参数给定风险值r，r值一般用小数表示。而$p=(1-r)$可称作保险值。给定风险值在目前还没有一套完善的方法，一种方法是由地质人员凭经验人为确定；另一种方法是根据含油气地质条件相似的邻区资料，通过统计分析确定。

例如，某个地质圈闭经过分析后给出如下风险值，见表3-2-8。

按（3-2-44）式计算便可以得到这个地质圈闭的风险系数R，即

表3-2-8 单一地质圈闭的风险数据表

地质参数	风险值（r）	保险值（1－r）
含油面积（S）	0.0	1.0
储集层厚度（H）	0.5	0.5
储集层孔隙度（Φ）	0.0	1.0
石油充满系数（D）	0.3	0.7
采收率（W）	0.0	1.0

$$
\begin{aligned}
R &= 1-\prod_{i=1}^{5}(1-r_i)\\
&= 1-(1.0\times0.5\times1.0\times0.7\times1.0)\\
&= 1-0.35\\
&= 0.65
\end{aligned}
$$

而保险值为0.35。

如果这个地质圈闭的石油储量分布函数为$AF(q_i)$，经过风险分析后，分布函数的每个值都要下降到原来概率35%的地方，见图3-2-13。

图3-2-13中的曲线①为风险分析前的石油储量分布函数；曲线②为风险分析后的石油储量分布函数。

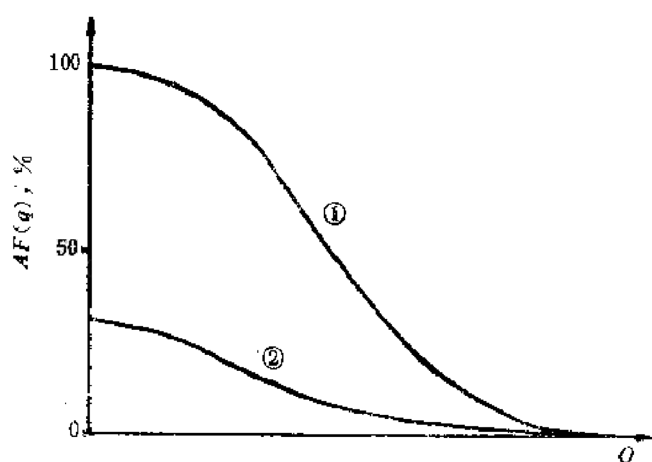


图3-2-13 风险分析后的石油储量分布函数

如果有一组在含油气地质条件上相类似的可能含油地质圈闭，总共10个。这组地质圈闭如按（3-2-45）式估算石油储量时，经过地质人员认真分析后，其各项地质参数的风险值如表3-2-9所示。

试问在这一组10个地质圈闭中，至少获得一个油藏的可能性有多大？这里有两种计算方法。

表3-2-9 一组可能含油圈闭的风险数据表

地质参数 \ (1－r) \ 圈闭序号	1	2	3	4	5	6	7	8	9	10
S	1.0	1.0	0.5	1.0	0.5	1.0	1.0	1.0	1.0	1.0
H	0.5	1.0	1.0	1.0	1.0	0.5	1.0	0.5	1.0	1.0
Φ	1.0	0.5	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5
D	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
W	1.0	1.0	1.0	0.5	1.0	1.0	0.5	1.0	0.5	1.0

（1）单独计算每个地质圈闭的风险系数，再计算一组地质圈闭的风险系数R

$$R=\prod_{j=1}^{10}\left[1-\prod_{i=1}^{5}(1-r_{ji})\right]=(1-0.25)^{10}=0.0563$$

最后再求出在这组地质圈闭中发现一个油藏的保险系数p。

$$p=1-R=1-0.0563=0.9437$$

即，发现一个油藏的可能性为94.37%。也就是说，发现一个油藏的可能性相当大。

（2）以最不利的地质因素，即以充满系数D的风险值来计算发现一个油藏的风险系数R

$$R=\prod_{i=1}^{10}r_i=(1-0.5)^{10}=0.00098$$

最后再求出在这组地质圈闭中发现一个油藏的保险系数p。

$$p=(1-0.00098)\times0.5=0.4951$$

即发现一个油藏的可能性为49.51%。也就是说，发现或不发现一个油藏的可能性大致各占一半。

通过计算，可见这两种计算方法给出的结果并不一样。为什么用第二种算法得到的保险系数偏小呢？其原因是在这5个地质因素中，每个圈闭的充满系数都存在风险，可见对于这组圈闭来说，充满系数的风险最大，所以其保险系数最小。

至于什么情况下用第一种算法，什么情况下用第二种算法，要根据探区的地质情况由地质人员选定。

风险分析时，如果地质参数间具有多层次结构，则要计算复合地质风险值。例如，某探区的地质风险决定于生油条件和储油条件。而生油条件与生油层厚度及生油相带有关，储油条件与储集层厚度及储集层相带有关，见表3-2-10。

表3-2-10　复合地质风险数据表

基础地质因素	风险值，r_{ji}	保险值，$(1-r_{ji})$	组合地质因素	风险值，r_j	保险值，$(1-r_j)$
生油层厚度	0.4	0.6	生油条件	0.4	0.6
生油层相带	0.0	1.0			
储集层厚度	0.1	0.9	储油条件	0.37	0.63
储集层相带	0.3	0.7			

复合地质风险可按下面的（3-2-46）式计算：

$$R=1-\prod_{j=1}^{m}\left\{1-\left[1-\prod_{i=1}^{n}(1-r_{ji})\right]\right\}$$

$$=1-\prod_{j=1}^{m}\prod_{i=1}^{n}(1-r_{ji})\qquad(3-2-46)$$

式中　R——复合风险值；

r_{ji}——第j项组合地质因素的第i个基础地质因素的风险值。

表3-2-10中的地质风险数据，按（3-2-46）式计算，其复合地质风险值如下：

$$R=1-(0.6\times0.1)\times(0.9\times0.7)$$
$$=1-0.378$$
$$=0.622$$

如果地质数据间的结构层次不止两层，则复合地质风险值可按（3-2-47）式计算：

$$R=1-\prod_{j_1=1}^{m_1}\prod_{j_2=1}^{m_2}\cdots\prod_{i=1}^{n}(1-r_{j_1j_2\cdots i}) \qquad (3\text{-}2\text{-}47)$$

从以上计算中可以看出，求逆概率是风险分析的主要算法。

七、风险分析后的石油资源量求和计算

当每个局部含油地质单元的石油资源量都经过风险分析后，求含油区的石油资源总量时，做法是：用随机数γ_s对第j个局部含油地质单元进行抽样计算，若随机数的值大于保险系数（$1-r$），由于入口值不能与分布函数$AF(q_j)$相交，该次对第j个局部含油地质单元的抽样结果Q_{jw}应等于0；只有在随机数的值小于或等于保险系数（$1-r$）时，才能通过插值计算得到抽样结果Q_{jw}，见图3-2-14。

从图3-2-14可以看出，用〔0，1〕区间上的第s个随机数γ_s对第j个局部含油地质单元抽样时，若$\gamma_s>(1-r)$，应令$Q_{jw}=0$。由于含油区共有m个局部含油地质单元，下次再对第j个局部含油地质单元进行抽样时，所用的随机数应为序列中的第$s+m$个，若$\gamma_{s+m}<(1+r)$，则可由插值计算得到Q_{jw+1}。

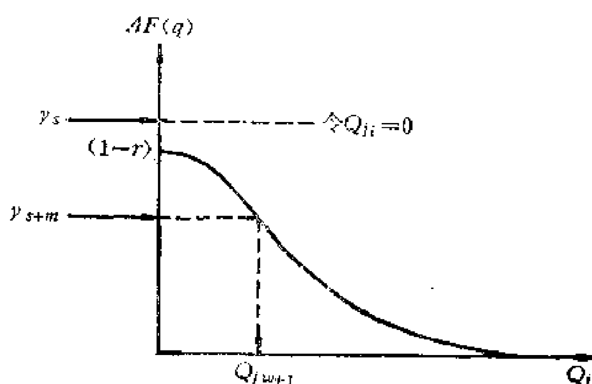


图3-2-14 风险分析后的抽样计算

图3-2-15是风险分析后，由局部含油地质单元求含油区石油资源总量的示意图。

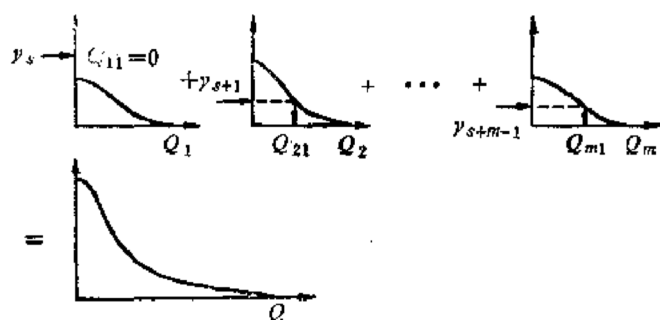


图3-2-15 风险分析后计算石油资源总量的示意图

由图3-2-15可知，假如第s个随机数γ_s大于第一个局部含油地质单元的保险系数（$1-r_1$），则第一次抽样值Q_{11}应等于0；而第$s+1$个随机数γ_{s+1}小于第二个局部含油地质单元的保险系数（$1-r_2$），则第二个局部含油地质单元的第一次抽样值Q_{21}不等于0；……；直到用第$s+m-1$个随机数γ_{s+m-1}对第m个局部含油地质单元进行第一次抽样，而得到Q_{m1}。

将这m个抽样值累加，则可得到含油区石油资源总量的一个随机估计值Q_w，即

$$Q_v = \sum_{j=1}^{m} Q_{iv}$$

请注意，上式中的Q_{iv}有时为0。

如果一共抽样计算g次，则可得到含油区石油资源总量的g个随机估计值。最后用频率统计法可以求出含油区石油资源总量Q的分布函数$AF(q)$。

由风险分析后的局部含油地质单元的分布函数$AF(q_i)$求得的含油区石油资源总量的分布函数$AF(q)$，其曲线形态多呈偏态分布。出现这种现象的原因前面已多次提到，是由于风险分析后，当随机数的值大于（$1-r$）时，使得局部含油地质单元的一些抽样值为0之故。同时也造成许多抽样和的值偏小，因而使含油区石油资源总量区间小值一侧抽样和的频率增大，而向大值一侧抽样和的频率迅速变小。所以分布函数曲线的高峰偏向小值一侧，而大值一侧曲线缓慢下降，即呈现偏态的长尾分布。见图3-2-15。

当然，只是在被累加的局部含油地质单元的数量不太多的情况下，才会出现偏态的长尾分布。当被累加的局部含油地质单元的数量充分大时，含油区石油资源总量的分布函数必将按中心极限定理趋向正态分布。

八、蒙特卡罗法估算石油资源量的优点

许多探区的实践证明，用蒙特卡罗法估算石油资源量的效果很好。这种方法的优点如下：

（1）蒙特卡罗法给出的石油资源量估计值带有成功的把握性，即可以给出不同概率下的石油资源量估计值。

（2）蒙特卡罗法适用于任何形式的石油资源或石油储量的计算。如果公式中存在着除法关系，可以用其倒数形式变为乘法关系。

（3）抽样计算方法可以压制原始数据中个别离群数据的干扰作用。因为离群数据出现的概率很小，在大量的抽样模拟计算过程中干扰作用自然会降低。

（4）与其他概率统计法相比，蒙特卡罗法给出的石油资源量估计值区间较窄，因而更适于石油资源评价。

九、蒙特卡罗法的程序设计要点

一个计算石油资源量的蒙特卡罗法实用性计算程序，应当具备很全面的计算和输出功能，因而应该是一个功能很强的软件包。在程序设计时，应考虑如下几个方面：

（1）应当适用于任何形式的石油资源量计算公式，而在计算公式中可以包括任意个随机变量、地质常数或经验系数。

（2）应具备在原始数据量不同的情况下，构造随机变量分布函数的功能。因为在石油勘探阶段地质参数的数量可能少至几个，多至上千个，所以程序中应当有若干种（不少于5种）构造随机变量分布函数的子程序。

（3）应当能够计算任意个局部含油地质单元石油资源量的合计资源量，而合计资源量又可分为若干个级别。例如，计算一个探区的石油资源总量时，可能需要计算地质圈闭、二级构造带、地质凹陷、地质坳陷、沉积盆地的石油资源量，此时可分为5个级别，即有4级求

和运算。也就是说，程序不仅能计算单个局部含油地质单元的石油资源量，而且要具备多级求和计算功能。

（4）应当具备对石油资源量进行风险分析，以及风险分析后的资源量求和计算功能。

（5）计算程序应当具有很强的输出功能，可以在宽行打印机、绘图仪、屏幕上彩显输出随机变量的分布函数、局部含油地质单元资源量、各级合计资源量的分布函数曲线图形、数据表以及汇总数据表等等。

十、算　　例

我国某沉积盆地中的一个地质凹陷有三套生油层系。为估算该凹陷的远景石油资源量，可按氯仿沥青法计算。每套生油层系的石油资源量计算公式如下：

$$Q_i = SHDAK_1K_2 \tag{3-2-47}$$

式中　Q_i——每套生油层系的石油资源量；

S——生油岩分布面积；

H——生油岩厚度；

D——生油岩密度；

A——氯仿沥青含量；

K_1——排烃系数；

K_2——聚集系数。

全凹陷的石油资源量计算公式为

$$Q = \sum_{i=1}^{3} Q_i$$

三套生油层系的地质参数见表3-2-11。

表3-2-11　三套生油岩层系的地质参数表

地质参数 ＼ 层系		第一套层系	第二套层系	第三套层系
生油岩分布面积S（km^2）		14000	7000	3000
生油岩密度D（$10^8t/km^3$）		23	23	23
排烃系数K_1		0.44	0.48	0.43
聚集系数K_2		0.111	0.111	0.111
生油岩厚度 H（km）	数据个数	140	70	30
	取值范围	0.1～1.0	0.1～1.0	0.1～0.5
氯仿沥青含量A（%）	数据个数	37	37	21
	取值范围	0.03～1.74	0.02～2.08	0.03～1.70

在（3-2-47）式中，生油岩分布面积S、生油岩密度D为地质常数；排烃系数K_1、聚集系数K_2是由地质类比法确定的经验系数；而生油岩厚度H、氯仿沥青含量A则为有一定取

值范围的随机变量。

按频率统计法计算得到第一套生油层系厚度的密度分布及分布函数见图3-2-16及图3-2-17。

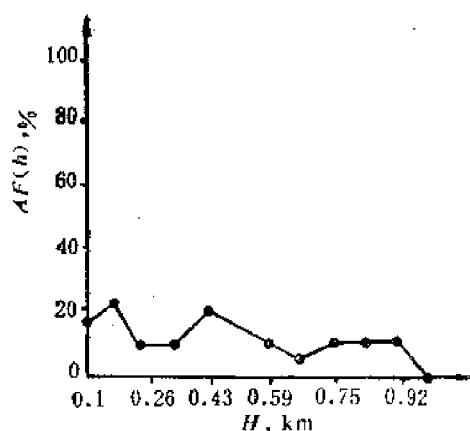


图3-2-16 第一套生油岩厚度的密度分布

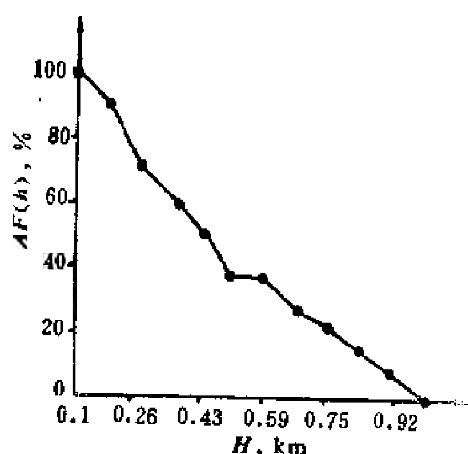


图3-2-17 第一套生油岩厚度的分布函数

第一套生油岩氯仿沥青含量的密度分布及分布函数见图3-2-18及图3-2-19。

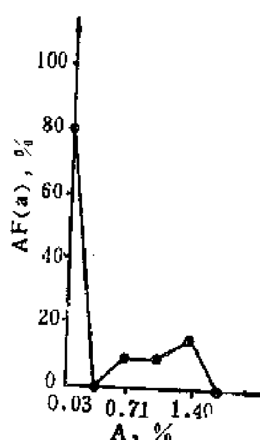


图3-2-18 第一套生油岩氯仿沥青含量的密度分布

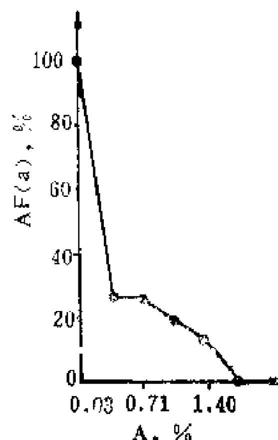


图3-2-19 第一套生油岩氯仿沥青含量的分布函数

计算后得到第一套生油岩层系的石油资源量Q_1的分布函数见图3-2-20。全凹陷三套生油层系的石油资源总量Q的分布函数见图3-2-21。

三套生油岩层系石油资源量Q_1、Q_2，Q_3及全凹陷石油资源总量Q在各概率下的数据汇总于表3-2-12中。

从表3-2-12中的数据可以发现，石油资源量Q_1、Q_2、Q_3累加后，分布函数的曲线形态确有向中间收缩的现象。当概率为100%时：

$$Q=2.9207(\times 10^8 t)$$

而 $Q_1+Q_2+Q_3=0.5358+0.3206+0.0988=0.9552(\times 10^8 t)$

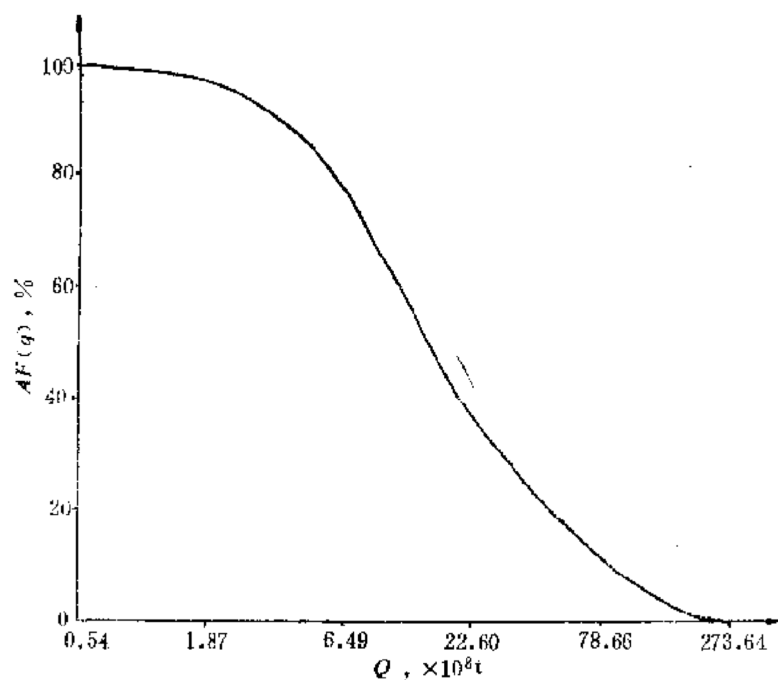


图3-2-20　第一套生油岩的石油资源量分布函数

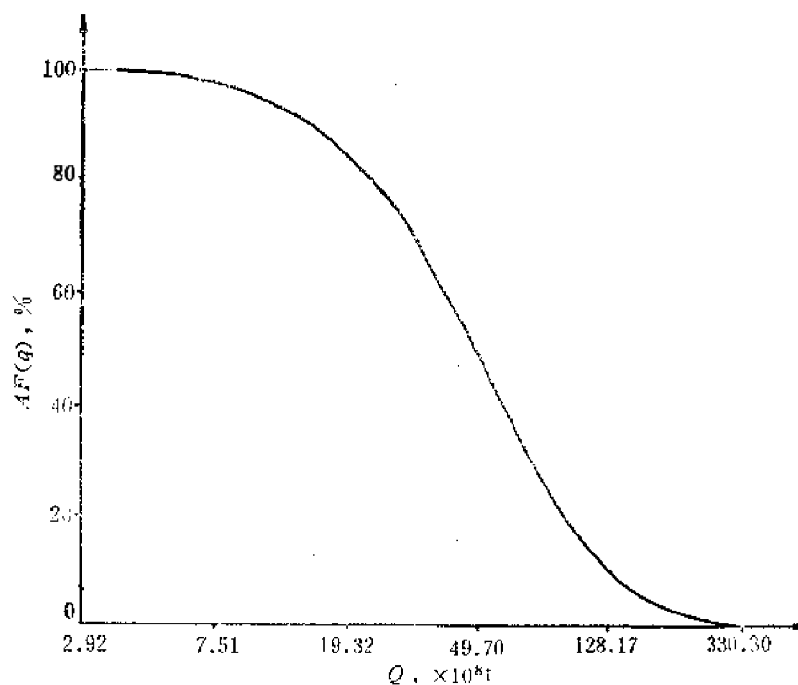


图3-2-21　全凹陷的石油资源量分布函数

可见　$Q>Q_1+Q_2+Q_3$

当概率为0%时：

$$Q=330.2972(\times 10^8 t)$$

而　$Q_1+Q_2+Q_3=273.6408+178.4241+27.8430=479.9079(\times 10^8 t)$

表3-2-12　全凹陷石油资源量汇总表

石油资源量(10^8t) 层位 / 概率(%)	第一套生油岩 Q_1	第二套生油岩 Q_2	第三套生油岩 Q_3	全凹陷 Q
100	0.5358	0.3206	0.0988	2.9207
95	2.5089	2.0563	0.2950	11.2233
90	3.6495	3.4874	0.4258	15.9204
85	4.7756	4.8753	0.5462	19.7932
80	5.8729	6.3479	0.6686	23.9546
75	6.8845	7.9966	0.7791	27.7634
70	8.1289	10.1081	0.8814	31.3018
65	9.3880	12.7875	1.0046	35.3488
60	11.2046	15.4970	1.1172	39.7212
55	13.2408	18.1056	1.2553	44.8798
50	15.5740	21.1254	1.4063	50.6544
45	18.4746	24.5984	1.5742	56.9070
40	22.1079	29.0008	1.8469	64.1746
35	25.5952	34.9261	2.1640	70.6543
30	31.0119	43.2521	2.4693	77.7917
25	36.9195	53.4797	2.8498	89.0505
20	46.6376	64.4440	3.3595	103.4409
15	60.5641	76.2474	4.2063	119.8079
10	99.1854	90.8671	6.6153	146.6928
5	148.8126	117.9727	11.5845	180.4146
0	273.6408	178.4241	27.8430	330.2972

可见　$Q<Q_1+Q_2+Q_3$

由于Q_1、Q_2、Q_3及Q的分布函数属于偏态分布（图3-2-20、图3-2-21中石油资源量坐标轴的区间是变换后的不等间距区间），大约在概率20%处，有

$$Q=103.4409(\times 10^8\mathrm{t})$$

$$Q_1+Q_2+Q_3=46.6376+64.4440+3.3595$$

$$=114.4411(\times 10^8\mathrm{t})$$

所以　$Q\approx Q_1+Q_2+Q_3$

第二节　翁（Weng）旋回模型

如果某一体系具有从兴起到衰亡的全过程，则这一过程可称作一个生命旋回。对于生命总

量有限的一些体系，例如对于非再生矿产资源的开采，可用Weng旋回模型进行描述和预测。

一、Weng旋回模型及其性质

客观世界中“从无到有”突然出现的实际体系，是一种不连续体系。如果时间从负到正，体系Q的不连续性可以表示为

$$Q=\begin{cases}0 & (t<0)\\ Q & (t>0)\end{cases}$$

如果体系的发展速度$\frac{dQ}{dt}$正比于实际存在的“现状”或“基础”Q，为了引入不连续过程，假设$\frac{dQ}{dt}$正比于$\left(\frac{x}{t}-1\right)$因子，其中$x$为$Q$达到顶峰的时间，这样可写出

$$\frac{dQ}{dt}=Q\left(\frac{x}{t}-1\right)$$

当t略大于0，如$t=\varepsilon$时，有

$$\frac{dQ}{dt}=\begin{cases}\left(\frac{dQ}{dt}\right) & \left(t=0, Q=0, \left(\frac{x}{t}-1\right)=\infty\right)\\ 0 & (t=x)\\ -Q & (t=\infty)\end{cases}$$

式中$\left(\frac{dQ}{dt}\right)_\varepsilon$是实际存在的非零有限常量。解上面的微分方程得到

$$\ln Q=x\ln t-t+\ln A$$

式中A为积分常量。上式可写成

$$Q=At^x e^{-t} \qquad (t\geqslant 0) \qquad (3\text{-}2\text{-}48)$$

（3-2-48）式就是Weng旋回模型。从该式可看出，体系Q的兴衰正比于一兴一衰两个因子。Q的兴起正比于时间t的x次方，即$Q\propto t^x$；Q的衰亡正比于时间t的负指数函数，即$Q\propto e^{-t}$。因此，体系Q是时间t的函数，而t是时间间隔与系数C的比值。所以，Weng旋回模型应表示为

$$\begin{cases}Q_t=At^x e^{-t}\\ t=(T-T_0)/C\end{cases} \qquad (t\geqslant 0) \qquad (3\text{-}2\text{-}49)$$

式中　x——某一正实数；

T_0——生命起始时刻；

T——生命过程中的某时刻；

A.C——拟合系数。

Weng旋回模型具有下列性质：

（1）$\frac{dQ_t}{dt}=Axt^{x-1}e^{-t}-At^x e^{-t}$

$$=Ax\frac{t^x}{t}e^{-t}-At^xe^{-t}$$

$$=At^xe^{-t}\left(\frac{x}{t}-1\right)$$

$$=Q_t\left(\frac{x}{t}-1\right)$$

所以，当$t<x$时，$\frac{dQ_t}{dt}>0$；

当$t=x$时，$\frac{dQ_t}{dt}=0$；

当$t>x$时，$\frac{dQ_t}{dt}<0$。

（2）$$\frac{d^2Q_t}{dt^2}=A[x(x-1)t^{x-2}e^{-t}-xt^{x-1}e^{-t}-xt^{x-1}e^{-t}+e^{-t}t^x]$$

$$=At^xe^{-t}\left[\frac{x(x-1)}{t^2}-\frac{2x}{t}+1\right]$$

$$=Q_t\left[\frac{x(x-1)-2xt+t^2}{t^2}\right]$$

$$=Q_t\left[\frac{1}{t^2}(x^2-x-2xt+t^2)\right]$$

$$=Q_t\frac{1}{t^2}[(t-x)^2-x]$$

所以，当$t=n+\sqrt{x}$时，$\frac{d^2Q_t}{dt^2}=0$；

当$t=n-\sqrt{x}$时，$\frac{d^2Q_t}{dt}=0$。

（3）当（3-3-4）式中的x为正整数，$t=\infty$时，对Q_t积分可得：

$$\int_0^\infty Q_t\,dt=A\Gamma(x+1)=Ax!=\Sigma_\infty Q_t$$

$\Sigma_\infty Q_t$可称作体系Q的生命总量。

（4）$$\frac{Q_t}{\Sigma_\infty Q_t}=\frac{t^xe^{-t}}{x!}$$

这一表达式和单项泊松分布在形式上相同。在许多数学文献和手册中，单项泊松分布常以λ代替t，一般表示平均值。随机型泊松分布的累计分布是对x的迭加，其中x是0，1，2，…，等正整数。

（5）确定型的Weng旋回，Q_t的累计式是对时间t的迭加，体系Q截至时间t的生命量可记为Σ_tQ_t，如果x也是正整数0，1，2，…，则可导出

$$\frac{\Sigma_t Q_t}{\Sigma_\infty Q_t}=\frac{\int_0^t Q_t \mathrm{d}t}{\int_0^\infty Q_t \mathrm{d}t}=\frac{1}{x!}\int_0^t t^x e^{-t}\mathrm{d}t$$

$$=\frac{1}{x!}\int_0^t t^x \mathrm{d}e^{-t}$$

$$=-\frac{t^x}{x!}e^{-t}+\frac{1}{x!}\int_0^t x t^{x-1}e^{-t}\mathrm{d}t$$

$$=-\frac{t^x}{x!}e^{-t}-\frac{t^{x-1}}{(x-1)!}e^{-t}-\cdots-\frac{t^2}{2!}e^{-t}+\int_0^t e^{-t}\mathrm{d}t$$

$$=-\frac{t^x}{x!}e^{-t}-\frac{t^{x-1}}{(x-1)!}e^{-t}-\cdots-\frac{t^2}{2!}e^{-t}-e^{-t}+1$$

$$=1-e^{-t}\sum_{i=0}^{x}\frac{t^i}{i!} \qquad (3\text{-}2\text{-}50)$$

鉴于上述，Weng旋回是个收敛模型，适用于生命总量有限体系的描述和预测。$t=0$时，(3-2-50)式等于0；$t=\infty$时，(3-2-50)式等于1。体系Q的发展过程中(3-2-50)式也是时间t的函数。

体系Q从兴起到衰亡大体可分为四个阶段，即

(1)加速上升阶段：$t=0\sim(x-\sqrt{x})$；

(2)一般上升阶段：$t=(x-\sqrt{x})\sim x$；

(3)一般下降阶段：$t=x\sim(x+\sqrt{x})$；

(4)缓慢下降阶段：$t=(x+\sqrt{x})\sim\infty$。

由Weng旋回模型的性质5可以导出

$$\Sigma_\infty Q_t=\frac{\Sigma_t Q_t}{1-e^{-t}\sum_{i=0}^{x}\frac{t^i}{i!}} \qquad (3\text{-}2\text{-}51)$$

对于生命总量有限体系，$\Sigma_t Q_t$的值可以通过实际观测获得，因而通过(3-2-51)式可以预测出体系Q的生命总量。对于非再生的石油资源，生命总量$\Sigma_\infty Q_t$就是一个油田的最终可采储量。因此，(3-2-51)式是预测石油储量的一个重要公式。

二、油田产量及最终可采储量的预测

油气田的形成是石油地质历史演变的结果，油气田中的石油、天然气是有限资源。油气田一经投入开发就成为一个体系，从油气田投产到产量枯竭是一个生命旋回。

用(3-2-49)式预测油气田产量的未来变化及最终可采储量时，式中的x为0，1，2，…等正整数；T_0为油气田的投产年份；T为油气田投产后的开采年份；A、C为表示油气田地质特征及开采方式的系数。

为确定(3-2-49)式中的拟合系数A，在实际计算时可作如下考虑，即当油气田的m个已知的逐年实际产量Q_t与(3-2-49)式中的$t^x e^{-t}$之间的相关系数最大时，认定x及C的值为

最佳值，此时可求拟合系数A，令

$$S=\sum_{i=0}^{m}(Q_i-Q_t)^2=\sum_{i=0}^{m}(Q_i-At^x e^{-t})^2$$

$$\frac{dS}{dA}=2\sum_{i=0}^{m}(Q_i-At^x e^{-t})(-t^x e^{-t})=0$$

$$\sum_{i=0}^{m}[Q_i(-t^x e^{-t})-At^x e^{-t}(-t^x e^{-t})]=0$$

$$\sum_{i=0}^{m}[A(t^x e^{-t})^2-Q_i t^x e^{-t}]=0$$

$$A\sum_{i=0}^{m}(t^x e^{-t})^2-\sum_{i=0}^{m}Q_i(t^x e^{-t})=0$$

$$A=\frac{\sum_{i=0}^{m}Q_i(t^x e^{-t})}{\sum_{i=0}^{m}(t^x e^{-t})} \quad (3\text{-}2\text{-}52)$$

$$R=\frac{\sum_{i=0}^{m}[(t^x e^{-t})-\overline{t^x e^{-t}}](Q_i-\overline{Q_i})}{\sqrt{\sum_{i=0}^{m}[(t^x e^{-t})-\overline{t^x e^{-t}}]^2\sum_{i=0}^{m}(Q_i-\overline{Q_i})^2}} \quad (3\text{-}2\text{-}53)$$

（3-2-53）式中

$$\overline{Q_i}=\frac{1}{m}\sum_{i=0}^{m}Q_i$$

$$\overline{t^x e^{-t}}=\frac{1}{m}\sum_{i=0}^{m}(t^x e^{-t})$$

至此，可用迭代法求出拟合系数x、C。

需要指出的是，确定拟合系数x、C时，除了要考虑相关系数R值尽可能大以外，还要使Q_t与最近时期的油田实际产量Q_i（$i=m$，$m-1$，…）尽可能接近。对于非正规开采的油田尤其需要如此。也就是说，在拟合时要尽量考虑近期产量，而早期产量可较少考虑，这可以称作"厚今薄古"的拟合原则。

在预测天然气田的产量时，特别是预测一个大的天然气区或一个国家乃至全球的天然气产量时，Weng旋回模型中应增加一个常数项Q_0，即

$$\begin{cases}Q_t=Q_0+At^x e^{-t}\\ t=(T-T_0)/C\end{cases} \quad (t\geqslant 0) \quad (3\text{-}2\text{-}54)$$

上式中的常数项Q_0可能包括目前开采工艺水平下尚不能完全采出的非正规天然气、地下水中的部分溶解气，也可能包括还在继续生成的生物气。可见常数项Q_0为发散部分，因而（3-2-54）式已与生命总量有限体系的含义有出入。但Q_0值一般不大，而且从数学上考

虑，（3-2-54）式只是把（3-2-49）式从坐标原点（0，0）沿纵坐标平移一段距离，这段距离的长度等于Q_0。因此，虽然模型中增加了一个发散部分，但并不影响模型的使用。

三、算　　例

作者应用Weng旋回模型，曾经对国内外170多个油气田的年产量及最终可采储量进行过预测。计算结果表明，这些油气田的已知实际产量与Weng旋回模型的拟合值之间的相关系数绝大多数都大于0.9；而正规开发的油气田的相关系数都在0.95以上。

应用Weng旋回模型预测时，已知的实际采油年数应该大于或等于5，也就是说，原始数据点数$m \geqslant 5$。

罗马什金油田是苏联仅次于萨马特洛尔油田的第二大油田，位于鞑靼自治共和国东部，发现于1948年，1952年投入开发。储油层为泥盆系Д$_1$及Д$_0$层砂岩，油层有效厚度15m，埋藏深度1650～1850m；油田面积3800km^2，地质储量45×10^8t；设计采收率为53.1%，可采储量24×10^8t；油层孔隙度为15～20%，平均渗透率300～400mD（$1D=0.987\times10^{-12}m^2$）；原始地层压力为175atm。1956年至1974年期间的产量在苏联占第一位。1970年为产量高峰年，其年产量为8150×10^4t，8000×10^4t的年产量保持了6年。稳产期末综合含水率为47.2%，累计采油11.967×10^8t，采出程度为26.59%（采出可采储量的49.86%）。1976年油田进入下降阶段，1976年至1979年期间每年产量递减225～430×10^4t，年递减率为2.8～5.9%，到1979年底已累计采油14.898×10^8t，采出程度为33.11%，含水率在60%以上。

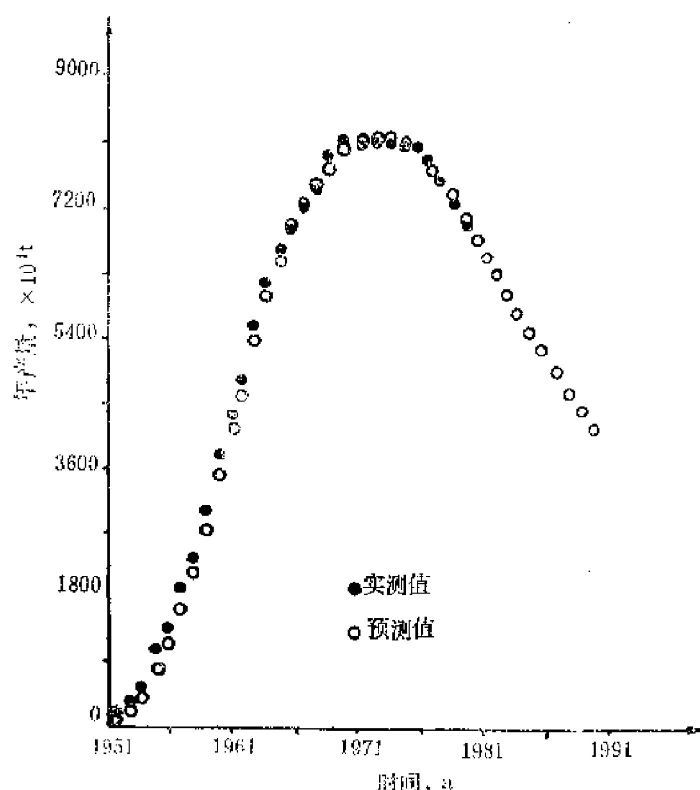


图3-2-22　罗马什金油田的产量预测图

经过计算得出罗马什金油田的Weng旋回模型表达式为

$$\begin{cases} Q_t = 6002.3t^3e^{-t} \\ t = (T-1951)/6.78 \end{cases}$$

按Weng旋回模型预测，罗马什金油田的最终可采储量$\Sigma_\infty Q_t = 25\times10^8$t。1952年至1979年期间的已知实际年产量与Weng旋回模型预测值之间的相关系数为0.99。

罗马什金油田的实际年产量以及用Weng旋回模型预测的年产量见表3-2-13及图3-2-22。

表3-2-13　罗马什金油田的产量预测表

年份	实际年产量（10^4t）	预测年产量（10^4t）	年份	实际年产量（10^4t）	预测年产量（10^4t）
1952	200	16.6	1972	8000	8056.1
1953	300	114.7	1973	8000	7992.4
1954	500	334.1	1974	8000	7880.2
1955	1000	683.3	1975	8000	7725.6
1956	1400	1151.5	1976	7775	7534.7
1957	1900	1716.9	1977	7500	7313.2
1958	2400	2352.5	1978	7230	7066.8
1959	3050	3030.1	1979	6800	6800.7
1960	3800	3722.7	1980		6519.6
1961	4400	4406.3	1981		6227.8
1962	5000	5060.5	1982		5929.2
1963	5600	5669.0	1983		5627.4
1964	6040	6219.2	1984		5325.3
1965	6600	6702.5	1985		5025.6
1966	6800	7113.3	1986		4730.4
1967	7000	7449.1	1987		4441.7
1968	7600	7709.6	1988		4160.9
1969	7900	7896.8	1989		3889.4
1970	8150	8013.8	1990		3628.0
1971	8000	8065.1			

四、小　结

（1）通过对国内外150多个油气田的试算结果表明，应用Weng旋回模型预测油气田的产量及最终可采储量是可行的。Weng旋回模型的重要意义在于，它能够根据油气田以往的实际产量预测出最终可采储量，特别是它可以发现不少油气田在原来计算储量时没有包括进去的潜在可采储量，而且这部分可采储量往往是相当可观的。例如前面提到的罗马什金油田，按Weng旋回模型预测的可采储量比原有的可采储量大约多1×10^8t。我国的一些大型油田，例如大庆油田也有类似情况。

（2）油气田的开发是生命总量有限体系的特例，可以认为Weng旋回模型将会在更多的领域中得到应用。例如，有人曾用此模型预测收音机的市场销量，也取得了满意效果。

（3）Weng旋回模型是一种唯象的基值预测模型。所谓唯象是指对信息的定义和性质不

作任何事先假设，而是从实际资料（例如油气田的生产记录）中找出信息，以这种拟合信息为基础的预测称为“基值预测”。对于和人类活动有关的许多体系，基值预测往往落后于人类的未来实践。原因是唯象拟合信息只能取得最后信息以前的信源状态，只能反映当前人类科学技术水平下的体系变化，而不可能反映出今后科学技术发展对预测体系的影响，例如油田开发采用了新工艺，改造低产油层，调整开发方案等等。所以预测仅是对未来的展望和分析，绝非最后结论。

此外，由于出发点不同，所站角度不同，假设条件不同，基础依据不同，应用方法不同等等，对同一问题的预测结果可能不一致。

第三节　油田规模序列法

“油田规模”（Qilfield Size）是指油气田的最终可采储量。如果某个含油气区经过详细勘探后，发现了全部油气田，并且查明了每个油气田的最终可采储量，那么，按最终可采储量由大到小进行排列，所得的顺序称为油田规模序列。

一、油田规模序列法的内涵

国内外许多含油气区的统计资料表明，当一个含油气区的一些油气田被发现后，如果以油田规模为纵坐标，以油田规模的序号为横坐标，在双对数坐标纸上展点作图大致可以得到一条直线，见图3-2-23。

根据这一规律，可以在探区的早期或中期勘探阶段，由已发现油气田的油田规模序列，预测尚未发现的油气田储量以及整个探区的油气总储量。这种预测方法称为油田规模序列法。

美国学者齐波夫（G.P.Zipf)于1949年在他所著的《人类行为与最小省力原则》一书中提出一种规律，这个规律可以表述如下：**将一组离散型随机变量，由大到小进行排列，如果最大的数值是第二大数值的两倍，是第三大数值的三倍，……，依此类推，则称这组离散型随机变量服从齐波夫定律。**

本世纪70年代以后，随着计算机技术的普及应用，齐波夫定律逐渐为人们所重视，1977年考兰德首先用齐波夫定律预测了赞比亚铜矿带的铜总量，其结果得到了在该地工作多年地质学家们的认可。1980年刘序琼在我国用齐波夫定律预测了一个铀矿床的资源。其后，人们应用齐波夫定律研究勘探地区的金属矿床或油气田的规模序列，借以预测尚未发现的金属矿产资源或油气资源。

实际上，齐波夫定律是巴内托（Pareto）于1927年所提出的定律的特例。巴内托定律可以表述为如下关系式：

$$\frac{Q_m}{Q_n}=\left(\frac{n}{m}\right)^k \qquad (3\text{-}2\text{-}55)$$

式中　Q_m——序号等于m的随机变量的数值；

Q_n——序号等于n的随机变量的数值；

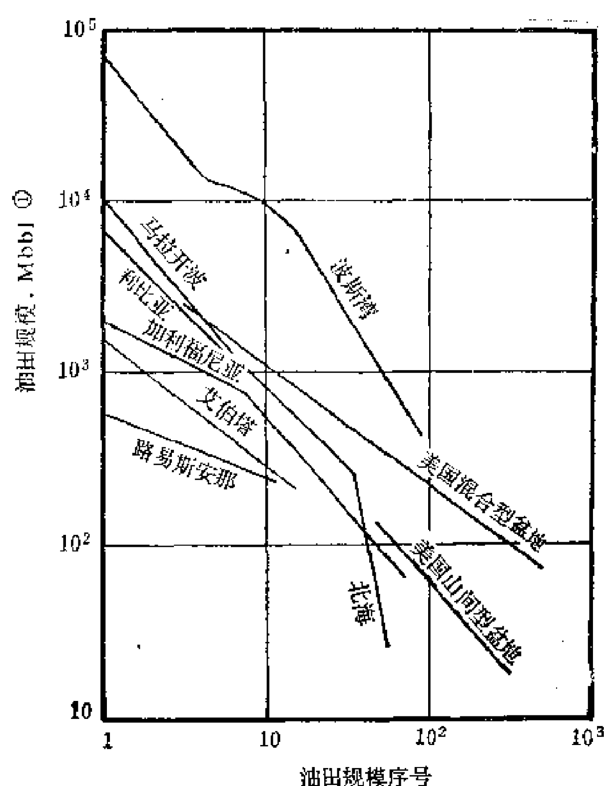


图3-2-23　世界主要含油气地区的油田规模序列

①1bbl＝159L

k——实数；

m、n——1，2，…整数序列中的任一数值，但$m\neq n$

当（3-2-55）式中的$k=1$时，则为齐波夫定律，即

$$\frac{Q_m}{Q_n}=\frac{n}{m} \tag{3-2-56}$$

或者　$mQ_m=nQ_n$

一个含油气地区内一组油气田的石油储量属于离散型随机变量，当最大的（第一号）油田被发现后，其石油储量为Q_{max}，若油田规模序列符合齐波夫定律时，则有

$$Q_{max}=nQ_n \tag{3-2-57}$$

而
$$Q_n=\frac{Q_{max}}{n}$$

假如含油气区总共有t个油气田，则全区的石油总储量SQ为

$$SQ=\sum_{i=1}^{t}\left(\frac{Q_{max}}{i}\right) \tag{3-2-58}$$

对（3-2-56）式的两边取对数，则有

$$\lg\left(\frac{Q_m}{Q_n}\right)=\lg\left(\frac{n}{m}\right)$$

$$\lg Q_m-\lg Q_n=-(\lg m-\lg n)$$

$$\frac{\lg Q_m-\lg Q_n}{\lg m-\lg n}=-1 \qquad (3\text{-}2\text{-}59)$$

因而，在双对数坐标纸上，以油田的石油储量Q_i为纵坐标，以油田的序号i为横坐标作图，则数据点的连线为斜率等于－1的直线。以上的（3-2-56）、（3-2-57）、（3-2-58)、（3-2-59）式就是目前国内外有些人所说的预测矿产资源或油气资源的齐波夫定律的不同表达形式。

但是，从图3-2-23可见，世界上主要含油气区的多数地区并不符合齐波夫定律，而是符合适应范围更广的巴内托定律。

对（3-2-55）式两边取对数，则有

$$\lg\left(\frac{Q_m}{Q_n}\right)=\lg\left(\frac{n}{m}\right)^k$$

$$\frac{\lg Q_m-\lg Q_n}{\lg m-\lg n}=-k \qquad (3\text{-}2\text{-}60)$$

因而在双对数坐标纸上作图，则数据点的连线为斜率等于－k的直线，这样便与图3-2-23中的所有含油气地区的统计规律相符合了。所以，应当认为油田规模序列的分布规律服从巴内托定律；而齐波夫定律仅是巴内托定律的特例。

二、油田规模序列法的使用条件

油田规模序列法的实质是根据已发现的油气田储量，应用巴内托定律预测一个含油气地区中尚未发现的油气田储量（或资源量）以及全区总的石油储量（或资源量）的一种外推预测方法。

虽然世界上多数含油气地区的油田规模序列在一定程度上符合巴内托定律，然而，直至目前为止尚不能从油气形成的地质理论上圆满地解释油田规模序列的地质成因。但是，许多事实说明，任何地质过程都受概率法则支配，所以对于一个含油气地区的油田规模序列形成的原因，暂切可以从统计规律方面去理解。

油田规模序列法适用于一个完整的、独立的石油地质体系。所谓一个完整的、独立的石油地质体系是指该地质体系内的油气生成、运移、聚集以及其后的地质变迁都是在同一石油地质演化历史条件下发生的；或者说，目前所要预测的含油气地区中的油气田（或油气藏）的分布规律具有统一的形成原因。

根据国内外主要含油气地区的统计资料，（3-2-60）式中系数k值的变化范围在0.5至2.0之间。这一情况说明，石油地质问题的复杂性导致了油田规模序列分布的多样性。而系数k等于－1的齐波夫定律只是多种油田规模序列分布的特殊情况。

当一个大的含油气地区具有多期成油过程时，可能存在多个油田规模序列。在这种情况

下需要对多个序列的复合总体进行筛分，分解出成因不同或成油期不同的多个相互独立的油田规模序列。

三、油田规模序列法的计算过程

（1）油田规模序列的系数k可由熟悉含油气区情况的地质家们商定。一般可以借鉴与本含油气地区在地质条件上相似的含油气地区的资料。如果确定系数k有困难，可令$k=-\mathrm{tg}\theta$，而θ的角度值应限定在115°～155°范围内，并把这一区间分为若干个子区间，进行多次油田规模序列的拟合计算。例如，取角度值步长为5°时，则有如下9个区间间隔值：

$-\mathrm{tg}115^\circ=2.1445$；　$-\mathrm{tg}120^\circ=1.7321$；　$-\mathrm{tg}125^\circ=1.4281$；
$-\mathrm{tg}130^\circ=1.1918$；　$-\mathrm{tg}135^\circ=1.0000$；　$-\mathrm{tg}140^\circ=0.8391$；
$-\mathrm{tg}145^\circ=0.7002$；　$-\mathrm{tg}150^\circ=0.5774$；　$-\mathrm{tg}155^\circ=0.4663$。

其中$-\mathrm{tg}135^\circ=1$时为齐波夫定律。

（2）把探区中已发现的t个油田，按储量Q_i（$i=1, 2, \cdots, t$）由大到小进行排列，选择最大的油田储量Q_1作为推算点。

（3）如果探区中已发现的t个油田储量为$Q_1, Q_2, \cdots, Q_t$，则以推算点Q_1除Q_i，并求出其值的k次方根，得到如下序列A_i，即

$$A_i=\sqrt[k]{\frac{Q_i}{Q_1}} \qquad (i=1, 2, \cdots, t) \qquad (3\text{-}2\text{-}61)$$

（4）序列A_i中的每个元素，乘以某一正整数$n_i=1, 2, \cdots$，令其乘积值$A_in_i=b_i$，当所有已发现的t个油田的b_i值最大限度地接近下面矩阵某一行号h时记入下面的矩阵B中，即

$$B=\begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1t} \\ b_{21} & b_{22} & \cdots & b_{2t} \\ \cdots & \cdots & \cdots & \cdots \\ b_{m1} & b_{m2} & \cdots & b_{mt} \end{pmatrix} \begin{matrix} b_{1i}\approx 1 \\ b_{2i}\approx 2 \\ \cdots \\ b_{mi}\approx m \end{matrix}$$

计算矩阵中各行的标准差σ_h：

$$\sigma_h=\sqrt{\frac{1}{t}\sum_{i=1}^{t}(b_i-\bar{b})^2} \qquad (h=1, 2, \cdots, m)$$

$$\bar{b}=\frac{1}{t}\sum_{i=1}^{t}b_i$$

当矩阵中第m行的标准差σ_m小于给定误差EP时，即$\sigma_m<EP$（一般情况下可令$EP=0.01$～0.05），此时有：

$$b_{mi}=A_in_i=\sqrt[k]{\frac{Q_i}{Q_1}}\cdot n_i\approx m$$

即有 $$\frac{Q_i}{Q_1}\approx\left(\frac{m}{n_i}\right)^k$$

由于此时$A_i n_i$接近正整数m，所以在给定的误差范围内已符合巴内托定律，因此可以把矩阵B的第m行作为含油气区油田规模序列的预测模型。

（5）把预测模型序列b_{mi}中的每个元素除以A_i，则可得到含油气区中已发现油田储量Q_1，Q_2，…，Q_t在预测的油田规模序列中的序号n_i（秩），即

$$n_i=\frac{b_{mi}}{A_i}\qquad(i=1,\ 2,\ \cdots,\ t)\tag{3-2-62}$$

式中 n_i——已发现的第i个油田的储量在预测的油田规模序列中的序号；

b_{mi}——矩阵B中第m行的第i列元素；

m——矩阵B的m行号，是已发现的最大油田储量Q_1在预测的油田规模序列中的序号（秩），$m\geqslant 1$。

（6）含油气区中已发现的任何一个油田储量Q_i（$i=1,\ 2\cdots,\ t$），乘以预测序号n_i的k次方幂，则为预测的最大（第一号）油田储量$\hat{Q}_{i\max}$。如果所有已发现油田的储量都是可靠的，则应以所有已发现油田的储量推算$\hat{Q}_{i\max}$的平均值，作为含油气区中预测的最大油田储量，即

$$\hat{Q}_{\max}=\frac{1}{t}\sum_{i=1}^{t}Q_i n_i^k\tag{3-2-63}$$

（7）用预测的最大油田储量$\hat{Q}_{\max}$除以1^k，2^k，…，则得到探区中预测的油田规模序列$\hat{Q}_j$，即

$$\hat{Q}_j=\frac{\hat{Q}_{\max}}{j^k}\qquad(j=1,\ 2,\ \cdots,\ p)\tag{3-2-64}$$

当预测的油田规模序列中第$p+1$个储量值$\hat{Q}_{p+1}<Q_{\min}$时，可以截断预测序列。$Q_{\min}$为人为规定的在当时经济技术水平下最小经济油田的储量值。

（8）预测全探区总的石油储量（或资源量）$S\hat{Q}$

$$S\hat{Q}=\sum_{j=1}^{p}\hat{Q}_j=\sum_{j=1}^{p}\left(\frac{\hat{Q}_{\max}}{j^k}\right)\tag{3-2-65}$$

（9）按$k=\text{tg}\,115^\circ\sim\text{tg}\,155^\circ$范围内的步长，分别计算$s$个预测的油田规模序列$\hat{Q}_r$之中与已发现油田对应的预测值$\hat{Q}_{ri}$，再计算每个序列中已发现油气田的实际储量$Q_i$与所预测的储量之间的标准差$\sigma_r$

$$\sigma_r=\sqrt{\frac{1}{t}\sum_{i=1}^{t}(Q_i-\hat{Q}_{ri})^2}\qquad(r=1,\ 2,\ \cdots,\ s)\tag{3-2-66}$$

式中 Q_i——含油气区中已发现的第i个油田的实际储量；

$\hat{Q}_{ri}$——第r个预测序列中，与已发现的第i个油田对应的预测值。

最后在s个预测序列中，选定σ_i的值为最小的序列作为预测的油田规模序列。

（10）上述的计算结果只是经过数学运算后得出的预测值，是否符合实际的地质情况，还需要由熟悉含油气区地质情况的地质学家们商榷。

四、算　例

某探区在地质构造上属于一个独立的地质凹陷，面积较小，经过勘探发现4个小油田，石油地质储量分别是149.143，61.567，34.375，27.277（$\times10^4$t）。

（1）由于该凹陷是个新探区，所以难以确定油田规模序列的系数k，故需要通过多次拟合计算才能确定k值。为了叙述上的方便，这里把第9步的计算结果 $k=-\text{tg}120°=1.7321$在此引用以作示范。

（2）把已发现的4个油田，按储量由大到小排列如下：

$$Q_1=149.143\ (\times10^4\text{t})\qquad Q_2=61.567\ (\times10^4\text{t})$$

$$Q_3=34.375\ (\times10^4\text{t})\qquad Q_4=27.277\ (\times10^4\text{t})$$

以其中最大的油田储量149.143（$\times10^4$t）作为推算点。

（3）用推算点Q_1去除Q_1，Q_2，Q_3，Q_4，并求所得之商的k次方根，得到序列A_i：

$$A_1=\sqrt[k]{\frac{Q_1}{Q_1}}=1.0\qquad A_2=\sqrt[k]{\frac{Q_2}{Q_1}}=0.6$$

$$A_3=\sqrt[k]{\frac{Q_3}{Q_1}}=0.4286\qquad A_4=\sqrt[k]{\frac{Q_4}{Q_1}}=0.375$$

（4）把序列A_i（$i=1,2,3,4$）乘以某一正整数，使其乘积值A_in_i最大限度地接近下面矩阵的行号，并记入下面矩阵B

$$B=\begin{bmatrix}1.0 & 1.2 & 0.8572 & 1.125\\ 2.0 & 1.8 & 2.143 & 1.875\\ 3.0 & 3.0 & 3.000 & 3.000\end{bmatrix}\quad\begin{matrix}b_{1i}\approx1\\ b_{2i}\approx2\\ b_{3i}\approx3\end{matrix}$$

矩阵B中的第1行各元素是由A_i与n_i相乘得到的，其中n_i为某一正整数，即

$$b_{11}=1.0\times1=1.0\qquad b_{12}=0.6\times2=1.2$$

$$b_{13}=0.4286\times2=0.8572\qquad b_{14}=0.375\times3=1.125$$

矩阵中第2行各元素为

$$b_{21}=1.0\times2=2.0\qquad b_{22}=0.6\times3=1.8$$

$$b_{23}=0.4286\times5=2.143\qquad b_{24}=0.375\times5=1.875$$

矩阵中第3行各元素为

$b_{31}=1.0\times3=3.0$　　　　$b_{32}=0.6\times5=3.0$

$b_{33}=0.4286\times7=3.000$　　　　$b_{34}=0.375\times8=3.000$

矩阵B计算到第3行时，标准差$\sigma_3=0.00063$，即可以认为已符合巴内托定律，因而可以把第3行作为油田序列规模的预测模型。

（5）预测模型序列b_{3i}中的每个元素除以A_i（$i=1, 2, 3, 4$），则得到已发现的4个油田储量Q_1，Q_2，Q_3，Q_4在预测的油田规模序列中的序号：

$$n_1=\frac{3.0}{1.0}=3 \qquad n_2=\frac{3.0}{0.6}=5 \qquad n_3=\frac{3.0}{0.4286}=7 \qquad n_4=\frac{3.0}{0.375}=8$$

即，已发现的4个油田储量Q_1，Q_2，Q_3，Q_4在预测的油田规模序列中，序号分别为3，5，7，8号。

（6）4个已发现油田的储量Q_1，Q_2，Q_3，Q_4分别乘以预测序号的k次方幂，即3^k，5^k，7^k，8^k，则得到预测的最大油田储量Q_{max}。

$$\hat{Q}_{1max}=149.143\times3^k=1000.0026$$

$$\hat{Q}_{2max}=61.567\times5^k=999.999$$

$$\hat{Q}_{3max}=34.375\times7^k=999.999$$

$$\hat{Q}_{4max}=27.277\times8^k=999.987$$

以这4个预测值的平均值1000×10^4，作为含油气区中预测的最大油田储量$\hat{Q}_{max}$。

（7）$\hat{Q}_{max}$分别除以1^k，2^k，…，则得到含油气区中预测的油田规模序列$\hat{Q}_1$，$\hat{Q}_2$，…，$\hat{Q}_p$。这里暂定最小经济油田的储量值$Q_{min}=10$（$\times10^4$t），则得到如下预测结果（单位为$\times10^4$t）：

$\hat{Q}_1=1000.000$　　$\hat{Q}_2=301.022$　　$\hat{Q}_3=149.142$（Q_1）

$\hat{Q}_4=90.615$　　$\hat{Q}_5=61.567$（Q_2）　　$\hat{Q}_6=44.895$

$\hat{Q}_7=34.375$（Q_3）　　$\hat{Q}_8=27.277$（Q_4）　　$\hat{Q}_9=22.243$

$\hat{Q}_{10}=18.533$　　$\hat{Q}_{11}=15.713$　　$\hat{Q}_{12}=13.513$

$\hat{Q}_{13}=11.765$　　$\hat{Q}_{14}=10.348$

用上述预测结果在双对数纸上展点作图，点的连线成一条直线，其斜率为$-k=-1.7321$，见图3-2-24。图中的黑点为已知油田，空圈为预测的油田。

（8）含油气区的石油储量总和SQ为

$$SQ=\sum_{i=1}^{14}\hat{Q}_i=1801.004\ (\times10^4\text{t})$$

（9）为预测该含油气区的油田规模序列，总共作了9次拟合计算。当$\theta=120^\circ$，即$-\text{tg}120^\circ=1.732$时，拟合效果最佳，已发现的4个油田储量与所预测的储量之间的标准差很小，$\sigma_r=0.00063$，所以被选定为预测序列。

（10）这一计算结果，经过熟悉含油气区地质情况的地质家们讨论，认为比较符合实际情况。

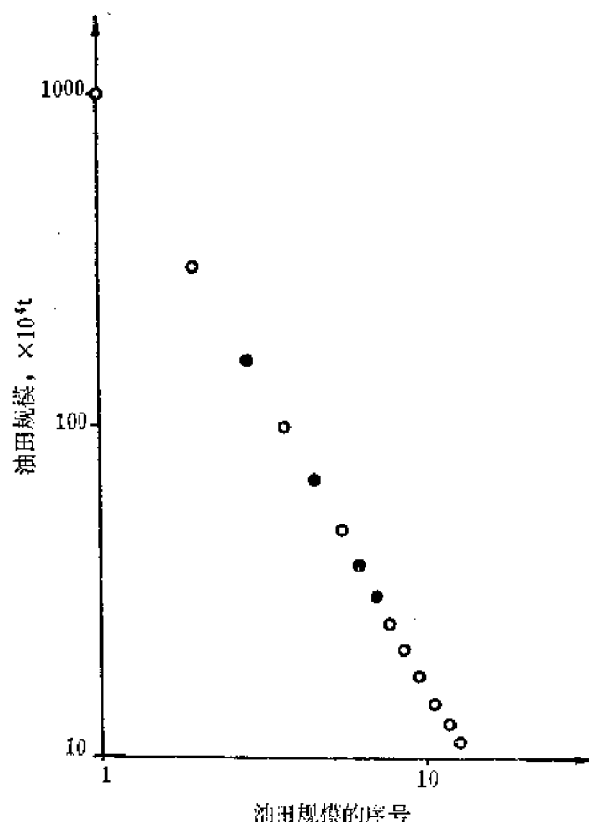


图3-2-24　某含油气区的油田规模序列

五、讨　　论

（1）近年来，国内不少地质研究单位已应用齐波夫定律预测各种金属矿床及油气田的储量。这里需要再次指出，齐波夫定律仅是巴内托定律的特例，所以，应当以巴内托定律作为金属矿床及油气田规模序列的理论预测模型。

（2）地质学的研究对象，包括各种地质过程及观测结果，它们普遍地受概率法则支配或影响。因而，许多人认为地质现象可视为随机事件，而地质观测数据具有随机变量性质。苏联的著名地质学者维斯捷列乌斯于1977年曾指出："地质现象是由一些单个单元联合起来的，这种联合遵循概率法则"。

尽管有时对油田规模序列法的预测结果尚不能从地质理论方面作出满意的解释，但是，应当看到世界上各主要含油气地区的油田规模序列都普遍地服从或接近服从巴内托定律。因而，应用油田规模序列法预测含油气区中尚未发现的石油储量（或资源量）是一种可行的方法。

（3）当含油气区已发现一批油气田时，应当从中选用可靠的油田储量数据作为预测依据；而储量数据不可靠的绝对不要使用。例如，某个油田还在打探边井，有可能增加新的储量或者油田的储量参数还需要验证，那么，这个油田当时的已知储量可能偏小，因而不能作为预测依据，否则将会得出错误的预测结果。

（4）对含油气区中已发现的油气田进行合理的油田单元划分，是油田规模序列法的重要环节。这里所说的油田与油藏的含义相同，也就是说它是一个完整的、独立的含油单元。

对于大型油田，不应以人为划分的采油矿区作为含油单元。

（5）油田规模序列法的预测结果带有多解性。在系数b不清楚的情况下，要经过多次拟合计算才能选出已发现油田储量Q_i与对应的预测值$\hat{Q}_i$之间标准差最小的油田规模序列，而且这个预测结果还要经对含油气区地质情况十分熟悉的地质人员的认可才行。

（6）世界上主要含油气盆地的实际资料（图3-2-23）说明，有些含油气盆地，例如波斯湾、北海等含油气区的油田规模序列在双对数坐标图上并不是直线，而是由两段以上的直线或曲线构成的折线。这意味着这些含油气地区可能存在多个互相独立的油田规模序列，此时，应当对已发现的油田序列进行数学筛分，分解出互相独立的油田规模序列。

第四节 干酪根降解法

有机物质随沉积物逐渐被埋藏在地下深处之后，在适当的条件下形成化学结构很复杂的干酪根，后者是带有许多官能团、分子量很大的具有三度空间结构的有机物质。其后，在适当的温度、压力、催化物质作用下，干酪根将逐步降解生成油气，这就是干酪根降解生成油气的基本原理，也是目前有机生油学说的主要论点。

由研究干酪根降解生油理论而获得声誉的法国蒂索等人（B.Tissot），于1969年首次把化学动力学理论应用到生油理论研究中，提出了干酪根降解生油的数学模型，把生油理论研究工作向前推进了一步。目前，这一理论已成为油气生成的主导理论。

一、干酪根降解生油的基本概念

1. 干酪根降解生油的两个阶段

蒂索将干酪根在温度和时间因素作用下，向油气转化的过程分为两个阶段，即

干酪根（X）→降解的中间产物（Y）→最终产物（U）。

这一演化过程的中间产物是液态烃（石油），最终产物是气态烃（天然气）。因此，整个干酪根降解过程可划分为生油及成气两个阶段。

2. 生油潜量与活化能分布

生油潜量是指干酪根降解成为油气的最大潜力，而不同类型的干酪根具有不同的生油潜量。

由于干酪根是结构十分复杂的、分子量很大的有机物质，因而其化学活化能不能用单一数值表示。由于干酪根中包含多种官能团以及其他杂原子，所以在分子结构中有多种类型的键合。周所周知，各种键合发生反应时的活化能是各不相同的，即使是同一种键合，由于相邻官能团不同，活化能也不相同。因此，研究干酪根的化学反应性能，不能只用单一活化能数值，而要用由多种不同的活化能数值构成一个活化能密度分布来表示。

干酪根的活化能分布，实际上是由各种类型键合活化能数值描述的离散型密度分布。然而，要测定每个单一类型键合活化能是有困难的，因此，我们常用具有不同活化能反应物质的数量来表示之。

根据蒂索等人在实验室对实际样品的测定结果，干酪根所含键合的活化能分布范围在几kcal/mol●到80kcal/mol之间。他们研究了Ⅰ、Ⅱ、Ⅲ型干酪根的生油潜量及活化能分布，以

●1cal=4.1840J

活化能的6个离散值代替密度分布，见表3-2-14和图3-2-25。

表3-2-14　三种类型干酪根的活化能分布及生油潜量

活化能		干酪根类型					
类型	平均值	Ⅰ型		Ⅱ型		Ⅲ型	
E_{1i}	kcal/mol	X_{io}	A_{1i}	X_{io}	A_{1i}	X_{io}	A_{1i}
E_{11}	10	0.024	4.75×10^{4}	0.022	1.27×10^{5}	0.023	5.2×10^{3}
E_{12}	30	0.064	3.04×10^{16}	0.034	7.47×10^{16}	0.053	4.20×10^{16}
E_{13}	50	0.136	2.28×10^{26}	0.251	1.48×10^{27}	0.072	4.33×10^{25}
E_{14}	60	0.152	3.98×10^{30}	0.152	5.52×10^{29}	0.091	1.97×10^{32}
E_{15}	70	0.347	4.47×10^{31}	0.116	2.04×10^{35}	0.049	1.20×10^{33}
E_{16}	80	0.172	1.10×10^{34}	0.120	3.80×10^{35}	0.027	7.56×10^{31}
$X_0=\sum_i X_{io}$		0.895		0.695		0.313	
Y_0		0.051		0.035		0.018	

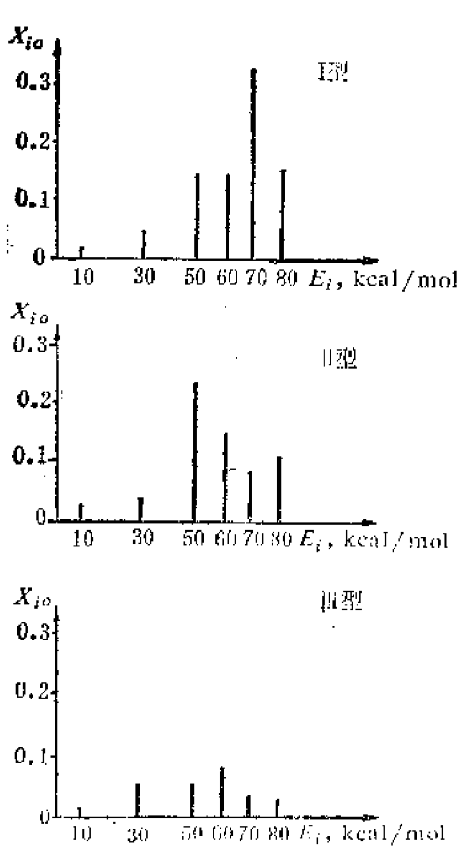


图3-2-25　干酪根活化能的密度分布图

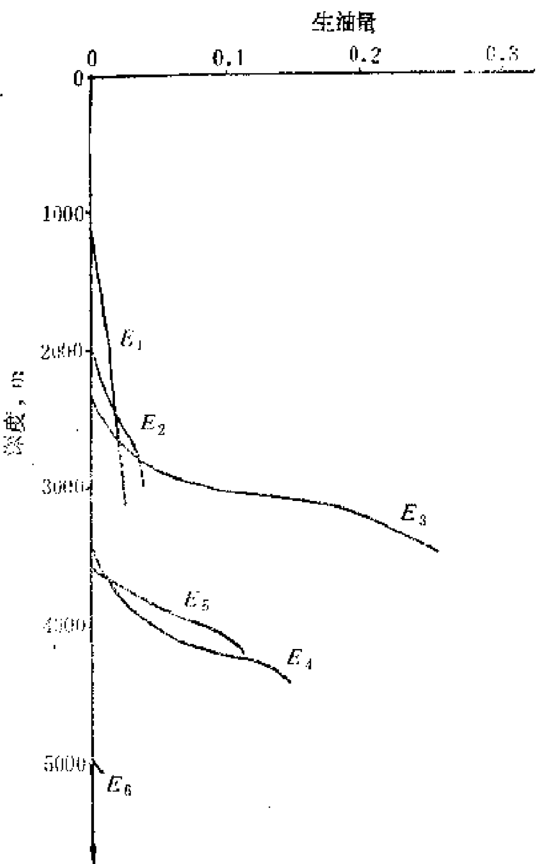


图3-2-26　Ⅱ型干酪根降解过程中6种活化能物质的降解曲线

表3-2-14中

E_{1i}——活化能，脚码1表示成油阶段，脚码i(i=1，2，…，6)表示6个活化能编号；

X_0——干酪根的生油潜量，是指单位重量干酪根能生成油气的最大比值，其中Ⅰ、Ⅱ、Ⅲ型干酪根的X_0值分别为0.895、0.695、0.313；

Y_0——积分初值，是指干酪根降解前已转化为烃类的比值，其中Ⅰ、Ⅱ、Ⅲ型干酪根的Y_0值分别为0.051、0.035、0.018。

上述情况说明，各种类型的干酪根具有不同的生油潜量，Ⅰ型最高，Ⅱ型居中，Ⅲ型最低；活化能的频率分布也不一样，Ⅰ型干酪根以活化能70kcal/mol为主，在单位重量的干酪根中占0.34；Ⅱ型干酪根以活化能 50kcal/mol 为主，占0.25；Ⅲ型干酪根以活化能60 kcal/mol为主，占0.091。

根据蒂索等人的上述数据，经过模拟计算可知，随着地层的温度增加，干酪根中活化能不同的6种物质是按活化能的增大顺序依次发生反应的。现以Ⅱ型干酪根为例，在图3-2-26中描绘了6种活化能物质随埋藏深度增加的变化情况，即描述了随着地温增加，干酪根中6种物质发生降解的情况，这说明引入活化能密度分布概念的重要性。

干酪根活化能的密度分布，从空间的观点来看，它表示了活化能不同的6种物质在干酪根中的相对比例关系；从时间的观点来看，它体现了活化能不同的6种物质在降解生油过程中的依次演变过程。就是说，随着埋藏深度的增加，地温不断升高，在这一过程中，干酪根中活化能最低的物质先发生反应，然后是活化能高的物质依次发生反应。亦即，由于活化能不同，发生降解反应的起始时间有先后之分。当几种物质都发生反应后，就同时存在几种平行反应，而各种物质反应的完成时间也是有先后之分。

3. 干酪根降解生油的数学模型

假定干酪根降解过程服从阿雷尼乌斯方程，则可建立如下微分方程组。

$$\begin{cases} \dfrac{dX_i}{dt}=-K_{1i}X_i \\ \dfrac{dU_1}{dt}=K_{21}Y \\ Y=\Sigma Y_i \\ \Sigma X_{io}+\Sigma Y_{io}+\Sigma U_{1o}=\Sigma X_i+\Sigma Y_i+\Sigma U_1 \end{cases} \tag{3-2-67}$$

$$\begin{cases} K_{1i}=A_{1i}\exp\left(-\dfrac{E_{1i}}{RT}\right) \\ K_{21}=A_{2i}\exp\left(-\dfrac{E_{21}}{RT}\right) \end{cases} \tag{3-2-68}$$

(3-2-67)式及(3-2-68)式中

t——时间(Ma)；

X_i——在时刻t，干酪根中第i种活化能物质的数量比例；

K——反应速率(Ma^{-1})；

A——频率因子(Ma^{-1})；

E——活化能(kcal/mol)；

K_{1i}——脚码1表示第一阶段(成油阶段)，脚码i表示第i种物质，K_{1i}表示成油阶段

干酪根中第i种活化能物质的反应速率；

A_{1i}——成油阶段干酪根中第i种活化能物质的频率因子；

E_{1i}——成油阶段干酪根中第i种物质的活化能；

K_{2j}——脚码2表示第二阶段（成气阶段），当最终降解产物为一种气体时$j=1$，K_{2j}表示成气阶段的反应速率；

A_{2j}——成气阶段的频率因子；

E_{2j}——成气阶段的活化能；

R——气体常数，$R=1.986$cal/mol；

T——绝对温度（K）；

Y——生油量（mkg碳/mkg有机碳）；

U_j——生气量（mkg碳/mkg有机碳）；

X_{i0}——时间$t=0$时，干酪根中第i种物质的初值；

Y_{i0}——时间$t=0$时，液态烃物质的数量；

U_{j0}——时间$t=0$时，气态烃物质的数量。

绝对温度T（K）由（3-2-69）式计算：

$$T=Gvt+T_0+273 \tag{3-2-69}$$

式中 G——地温梯度（°C/hm）

v——沉降速度（m/Ma）

t——时间（Ma）

T_0——地表年平均温度（°C）

上述数学模型是对生油机理的简要描述，尽管还不十分完善，如未考虑地层压力、催化条件等，但它毕竟是从机理上导出的理论模型。这一模型把时间、地温、生油量三者间的关系定量地联系起来，可以定量地给出油气生成数量。

需要指出，（3-2-67）方程组中前两式的等号右侧的反应速率K是随时间改变的，所以不能用简单的积分方法求解，而应采用数值积分方法求解。

二、油气生成量的计算步骤

1. 确定计算参数

（1）地质参数　包括生油岩地质时代及最大沉降深度（m）；沉降速度（m/Ma）；地温梯度（°C/hm）；地表年平均温度（°C）；生油岩分布面积（km^2）；生油岩厚度(m)；生油岩密度（$0.023\times10^8/km^2\cdot m$）。

（2）地球化学及热动力学参数　包括干酪含量（可用有机碳含量表示）；干酪根类型（计算时选用相应的活化能E及频率因子A）；干酪根生油潜量X_0及积分初值Y_0。

2. 求解数值积分

求解方程组（3-2-67）式及（3-2-68）式，可采用龙格—库塔法，求解第$n+1$步的积分公式如下：

$$Y_{n+1}=Y_n+\frac{1}{6}(k_1+2k_2+2k_3+k_4)$$

$$k_1=Hf(X_n,\ Y_n)$$

$$k_2=Hf\left(X_n+\frac{H}{2},\ Y_n+\frac{k_1}{2}\right)$$

$$k_3=Hf\left(X_n+\frac{H}{2},\ Y_n+\frac{k_2}{2}\right)$$

$$k_4=Hf(X_n+H,\ Y_n+k_3)$$

上面公式中的H为时间步长（Ma）；$f(X, Y)$为微分方程内函数式。

3. 计算过程

由于干酪根降解是由生油及生气两个阶段组成的，所以实际计算过程也分两步。从什么时间开始进行生气量计算是首先需要解决的问题。

确定生气阶段的起始时刻是一个比较复杂的问题，因此，应当综合考虑勘探地区的各种实际资料，例如镜煤反射率、气层在探区中实际出现的深度等资料来确定。实际计算时，应从干酪根降解过程中因升温而增加反应速度的基本原理来判断。即根据温度升高，反应速度增大的原则，如果温度继续增高时，生油速度反而减慢了，便认为增加的热能消耗于液态烃的裂解过程中了。因此，可将生油阶段中生油速度最大值所对应的时刻t作为生气阶段的起始时刻，从此点开始引用成气阶段的活化能E_2，频率因子A_2，进行生气量计算。

（1）单位生油量计算　以上所说的计算都是指干酪根的降解率（或称转化率），所以称作单位生油量。

令M时刻的生油量为YB，则有

$$YB(M)=X_0+Y_0-XI \tag{3-2-70}$$

式中　X_0——生油潜量，$X_0=\Sigma X_{io}$；

Y_0——原始液态烃数量，$Y_0=\Sigma Y_{io}$；

XI——为M时刻剩余的干酪根数量，由数值积分方法得到，$XI=\Sigma X_i$。

（2）单位生气量计算　在干酪根降解初期直至降解生油速度达到最大值之前，可以忽略不计非降解的其他成因生成的气体，例如生物成因的气体。这里的生气量是指从生气点开始后的降解生气量，即计算由液态烃（石油）裂解生成的气体数量。

进入生气阶段后，干酪根降解系统进入了三相状态，即有固体的干酪根、液态的石油及气态的天然气。令M时刻的生气量为$YG(L)$，则有

$$YG(L)=YB(M)-YU(L) \tag{3-2-71}$$

式中　$YG(L)$——生气量；

$YB(M)$——M时刻干酪根降解产物的总量；

$YU(L)$——降解三相系统中的液态烃数量。

（3）降解率计算　在M时刻，降解率KR为

$$KR(M)=\frac{X_0-\Sigma X_i}{X_0} \tag{3-2-72}$$

（4）生成速度计算　在M时刻，生成速度VB为

$$VB(M)=\frac{YB(M)-YB(M-1)}{HW} \tag{3-2-73}$$

式中　HW——时间增量，即为输出步长。

（5）生油岩的生油量计算　数值积分的结果可以给出单位干酪根的降解率，据此便能估算出勘探地区中某一生油岩分布地区的总生油量Q：

$$Q=SHDCX_0KR \tag{3-2-74}$$

式中　S——生油岩分布面积（km^2）；

H——生油岩厚度（m）；

D——生油岩密度，一般取$D=0.023\times10^8 t/km^2\cdot m$；

C——有机碳含量（%）；

X^0——干酪根生油潜量（包括生油潜量和原始液态烃）；

KR——降解率（%）。

三、算　　例

我国某小型陆相盆地，按上述方法估算其生油岩的总生油量。计算时借用了蒂索发表的热动力学参数，采用自动变步长方法进行积分计算。

1.　地质参数

（1）生油岩时代为第三纪；

（2）最大沉降深度$H=5000m$；

（3）沉降速度$V=100m/Ma$；

（4）地温梯度$G=3.6°C/hm$；

（5）地表年平均温度$T_0=14°C$；

（6）生油岩分布面积$S=347.4km^2$；

（7）生油岩密度$D=0.023\times10^8 t/km^2\cdot m$；

（8）生油岩厚度$H=409.5m$；

（9）有机碳含量$C=1.2\%$。

2.　地球化学及热动力学参数

（1）干酪根类型为Ⅱ型；

（2）生油阶段的X_{i0}、A_{1i}、X_0、Y_0选用了表3-2-14中的Ⅱ型干酪根的数据；

（3）生气阶段的活化能E_2，经过多次试算认为选用$E_2=70kcal/mol$，$A_2=0.2\times10^{36}$为宜。

3.　温度计算

在阿雷尼乌斯公式中，T（绝对温度）是e的负指数的分母组成部分，从有机质演化角度看，温度T是时间t的函数，即

$$T=Gvt+T_0+273$$

$$RT=RGvt+R(T_0+273)$$

对该地区来说：

$$RGv=7.150;$$

$$R(T_0+273)=569.982。$$

4.　计算结果

按上述参数计算，干酪根降解生油演化及降解率见图3-2-27及图3-2-28。

（1）干酪根降解生烃演化过程　从图3-2-27及图3-2-28可以看出，该地区干酪根降解

生油过程可分为三个阶段：

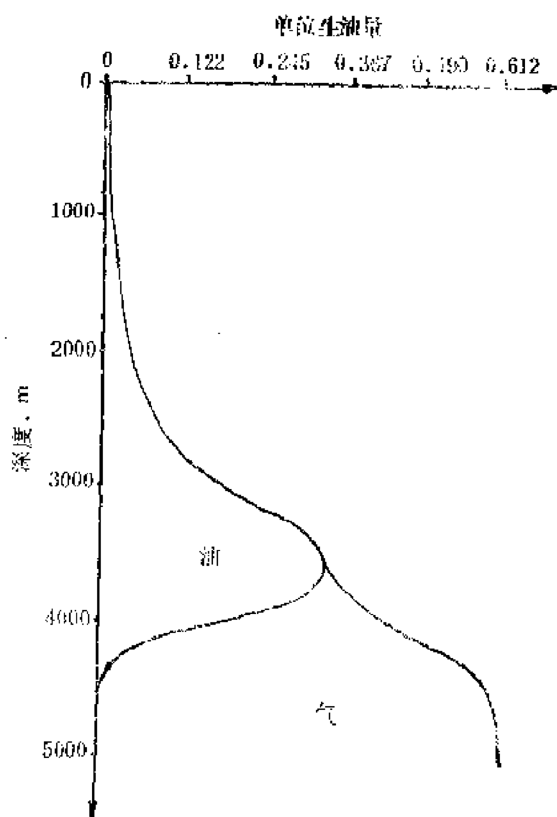


图3-2-27　干酪根降解生烃演化图

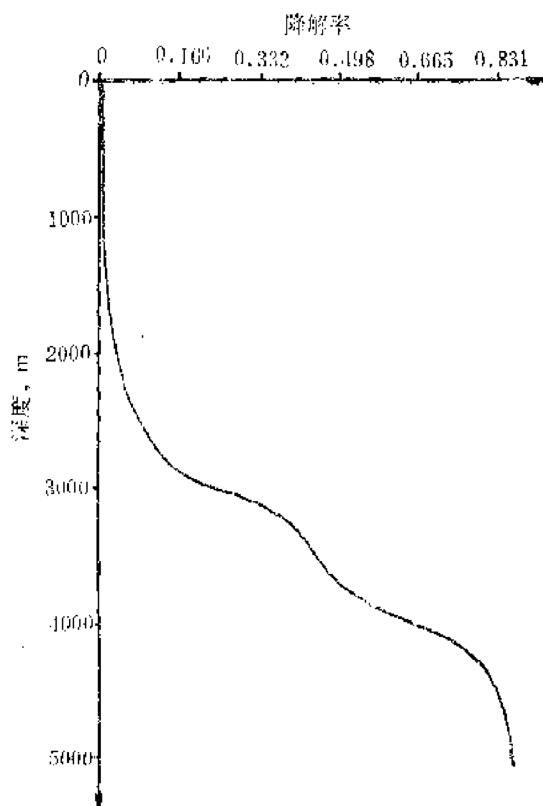


图3-2-28　干酪根降解率图

①初期生油阶段　从深度1500m开始至2200m，相当于地温由54°C升至79.2°C，生油速度由0.09%增加到0.37%。

②主要生油阶段　深度为2200m至3200m，地温增至115.2°C，生油速度达到5.2%。

③生气阶段　深度从3200m开始进入生气阶段，干酪根降解系统进入三相状态，液态烃（石油）逐渐减少。至4400m时，地温为158.4°C，干酪根降解产物全部为气，降解率达到82%。在深度4400m至5000m时，干酪根在高温条件下直接成气，降解率已达到83%。

（2）生油岩的总生油量　经过计算该地区的生油岩总生油量$Q=23.76\times10^{8}t$

最后需要指出，这个算例中借用蒂索发表的参数未必合理。

第五节　特尔菲法

特尔菲（Delphi）法目前的含义是对同一问题的多种见解或多种判断所进行综合处理的一种方法。在石油资源评价中，特尔菲法是一种客观地综合石油地质专家们的知识、经验或见解的技术。从集思广益的角度看，特尔菲法无疑是有价值的，特别是在探区的早期资源评价时更为有用。

用特尔菲法进行石油资源评价的前提要求是：已有的地质资料是够用的，参加石油资源

评价的地质专家是合格的。在这两个基本条件下，得出的评价结论才能被认为是可信的。

一、特尔菲法的要点

（1）要有一名负责人主持石油资源评价工作，这位负责人可称为特尔菲班长。由班长聘请若干名具有丰富经验及渊博知识的石油地质专家，组成勘探地区的资源评价小组。故此，也有人把特尔菲法称作专家评价法。

（2）资源评价小组中每个专家使用的评价方法，要由专家本人确定，特尔菲班长不能作任何干预，以期保证每个专家充分发挥自己的经验和才干。

（3）资源评价小组成员之间匿名，以防止因评价组内存在某些技术权威人士，而使评价意见产生倾向性。

（4）评价任务可以一次完成，也可以多次反复进行，即以匿名方式把前一轮每个专家的评价结果"反馈"给所有专家，让每个专家再次认真研究核对，并且重新给出下一轮的评价意见。

（5）每个专家的评价意见，最好以各种概率下资源量估计值的形式给出。而特尔菲班长则要根据各位专家的评价意见，综合出最终的评价结论，并以分布函数的形式表示，即应该给出不同置信水平下的石油资源量估计值。

二、特尔菲法的实施步骤

（1）确定勘探地区的评价范围，即各位专家进行资源评价的地域范围必须是同一的、唯一确定的。

（2）选择有经验的石油地质专家作为特尔菲班长。如果班长认为需要的话，也可选定若干名助手协助工作。班长负责主持整个石油资源评价工作；助手只对班长负责，作好班长委托的具体工作。

（3）由特尔菲班长负责选聘若干名石油地质专家组成一个石油资源评价小组。这些专家应当有丰富的实践经验，对评价地区的地质情况应当有较为详细的了解。特别是要求这些专家能够坚持实事求是的原则。

所聘请专家的数量，原则上越多越好，但特别要注意专家的质量。对于每位专家，班长可根据他以往的工作成就或在石油地质界的威望，给以不同的权，用以确定各位专家在评价组中的作用。当然，这个权值只能由班长自己掌握，而不宜公开，否则将会影响一些权值较小的专家情绪。

（4）由班长拟定向各位专家征询评价意见的表格以及征询内容。目前最常用的是一种以概率分布形式提问的征询表。例如，可向专家们提出如下一些问题：

①在评价区中发现10Mt以上石油资源量的可能性有多大？

②发现50Mt以上石油资源量的可能性有多大？

③发现100Mt以上石油资源量的可能性有多大？

④发现500Mt以上石油资源量的可能性有多大？

或者把问题反过来提出：

①在评价区至少能拿到多少石油资源量（即概率为100%时的资源量）？

②在概率为75%时的石油资源量有多少？

③在概率为50%时的石油资源量有多少？

④在概率为25%时的石油资源量有多少？

⑤最多可以拿到多少石油资源量（即概率为0%时的资源量）？

（5）向不熟悉或不习惯用概率方法估算石油资源量的专家们解释或说明用概率估值的意义和方法。一般情况下，多数专家乐于接受以概率形式的提问；个别专家因为不熟悉概率方法，坚持采用单一估值给出评价意见时，特尔菲班长也不必强求非用概率方法给出不行。因为只要经过简单的数学方法处理，单一的点估值也可以满足尔后的计算要求。但应当尽量使专家们不要采用单点估值方法。

（6）每个专家在详细了解评价区的地质情况后，根据自己的经验、知识、习惯，确定评价方法，并按特尔菲班长的征询内容回答问题。

（7）特尔菲班长将所有专家的评价结果，用约定的算法进行综合，得出对评价区的综合评价结论。

（8）由特尔菲班长与各位专家进行单线联系，讨论、商榷各位匿名专家的评价意见以及班长得出的综合评价结论。其目的是尽可能消除重大分歧，以求得尽可能的统一认识。

（9）各位专家根据特尔菲班长给出的第一轮评价结论，以及与班长的商榷结果，重新提出修改后的评价意见（或者坚持自己的原有评价意见）。然后回到第7步，再由班长综合出第二轮评价结论。如此反复，直到两轮评价结果无明显差别时，特尔菲班长即可认为完成了评价工作。

（10）以最终的评价结果，向上级领导或委托单位呈报评价区各种概率下石油资源量的估计值。

三、特尔菲班长的综合处理方法

虽然不少书籍、文献中都论述过特尔菲法的要点和步骤，但是，尚未见到有关特尔菲班长如何综合处理各位专家评价意见的具体算法。本书给出的两种综合算法是按特尔菲法的基本要点及工作步骤拟定的。

下面以一个算例来具体说明在石油资源评价中如何应用特尔菲法。

某沉积盆地经地震及少量钻井证实，该盆地为一个含油气远景区。为估算该盆地的石油资源量，评价单位选聘了特尔菲班长，由班长聘请了11名具有丰富经验的石油地质学家组成一个评价小组。特尔菲班长将自己掌握的全部地质资料印发给每个专家。每个专家在消化地质资料的基础上，用不同的找油理论及相应的预测方法，给出了评价意见，见表3-2-15。

为了叙述上的方便，这里约定如下：称第i个专家的石油资源量估计值为Q_i，对应的累积概率为AF_i（$i=1, 2, \cdots, R$）；若第i个专家给出了N个不同概率下石油资源量的估计值，则第i个专家的第j点估计值以Q_{ij}表示（$j=1, 2, \cdots, N$）。第i个专家估计值中的极小值与极大值分别以Q_{imin}及Q_{imax}表示。

前已述及，各位专家要尽量避免用单一的点估计值。但是，有些专家，例如表3-2-15中的6～9号专家，由于不习惯或不熟悉概率估值方法，他们坚持给出单一的点估计值。为了进行尔后的计算，需要作些处理。这里是采用小区间展开法进行处理，即认为这个点估计值相当于概率为50%时的估计值。

小区间展开方法是把这个点估计值除以某个数C（$C>1$），得到一个在数量上远小于估

表3-2-15　各位专家的评价意见汇总表

专家号	各位专家的权系数	资源量估值点数	各概率水平下的石油资源量估计值Q_{ij}（10^8t）				
R	W	N	100%	75%	50%	25%	0%
1	1	2	22.22				34.17
2	1	2	31.51				42.06
3	1	2	20.95				30.65
4	1	2	39.68				57.60
5	1	2	28.10				52.24
6	1	1			32.77		
7	1	1			46.00		
8	1	1			70.00		
9	1	1			65.00		
10	1	5	20.15	26.32	27.49	28.90	34.05
11	1	5	30.75	46.12	51.13	54.96	65.17

计值Q的Δq，即：

$$\Delta q=\frac{Q}{C} \tag{3-2-75}$$

再以Q为中心向两侧分别外推Δq，得到（$Q-\Delta q$）及（$Q+\Delta q$）两个值。这两个值可以看作是概率分别为100%及0%区间估计值的端点。

为了区别对待每个专家的作用，特尔菲班长可以用不公开的方式，赋给每个专家以不同的权系数W_i，一般取W_i为正整数。让经验丰富的专家具有较大的权，亦即认为一个有经验的专家所起的作用相当于两个或多个专家的作用。如果特尔菲班长对所有专家的评价意见一视同仁，则所有专家的权系数都赋值为1。在我们的实例中所有专家的权系数W_i均等于1。

有了以上准备，特尔菲班长即可以进行综合计算。本书给出两种综合算法，即概率加权法和加权抽样法。

1. 概率加权法

首先在R个专家所给出的所有石油资源量估计值Q_{ij}中，找出最大的估计值Q_{max}及最小的估计值Q_{min}。如果在R个专家中，有些专家的评价值是以点估计值给出的，则要在用小区间展开方法处理后，以（$Q_i-\Delta q_i$）及（$Q_i+\Delta q_i$）代替原来的点估计值Q_i，参加Q_{max}及Q_{min}的挑选。

选出的Q_{max}及Q_{min}作为最终评价结论的区间估计值的两个端点，进而再把这一区间分为m个子区间，并求出（$m+1$）个子区间的分隔值Q_k。

为求出（$m+1$）个与石油资源量Q_k相对应概率值的加权平均值AF_k，可按如下公式计算：

$$AF_k=\sum_{i=1}^{R}(AF_i\ W_i)/SW \qquad (k=1,\ 2,\ \cdots,\ m+1) \tag{3-2-76}$$

$$AF_i=\begin{cases}1 & (Q_k<Q_{imin}\text{时}) \\ \dfrac{(AF_{ij}-AF_{ij-1})(Q_k-Q_{ij-1})}{(Q_{ij}-Q_{ij-1})}+AF_{ij-1} & (Q_{imin}\leqslant Q_k\leqslant Q_{imax}\text{并且}Q_{ij-1}\leqslant Q_k\leqslant Q_{ij}) \\ 0 & (Q_k>Q_{imax}\text{时})\end{cases} \quad (3\text{-}2\text{-}77)$$

$(k=1, 2, \cdots, m+1)$ $(i=1, 2, \cdots, R)$ $(j=1, 2, \cdots, N)$

$$SW=\sum_{i=1}^{R}W_i \quad (3\text{-}2\text{-}78)$$

（3-2-76）、（3-2-77）、（3-2-78）式中

Q_k——第k个点的石油资源量估计值；

AF_k——Q_k的概率值；

AF_i——第i个专家Q_{ik}估计值的概率；

W_i——第i个专家的权系数；

SW——所有专家的权系数之和；

Q_{ij}——第i个专家的第j点资源量估计值；

AF_{ij}——Q_{ij}的概率值；

Q_{ij-1}——第i个专家的第（$j-1$）点资源量估计值；

AF_{ij-1}——Q_{ij-1}的概率值；

Q_{imax}——第i个专家最大的资源量估计值；

Q_{imin}——第i个专家最小的资源量估计值。

上面（3-2-77）式的含义是，当Q_k小于Q_{imin}时，令第i个专家估计值的概率AF_i为100%；当Q_k大于Q_{imax}时，令第i个专家估计值的概率AF_i为0%；当Q_k大于等于Q_{imin}同时小于等于Q_{imax}时，用线性插值求出第i个专家估计值的概率AF_i。

最后把计算出来的（$m+1$）个点（Q_k，AF_k）以分布函数的形式表示出来，这就是根据R个专家的评价意见综合出来的最终评价结论。这一结论给出了不同概率下的石油资源量的估计值。

需要指出的是，个别专家的离群估计值，是以小概率出现于最终的评价结论之中，对最终的评价结论影响不大。

2. 加权抽样法

加权抽样法是用随机数对R个专家给出的石油资源量分布函数，进行随机抽样计算。具体的作法是以每个专家的权系数W_i（W_i必须是正整数）作为抽样次数，再把R个专家的抽样值累加起来，除以R个专家的权系数之和SW，得到一个复合抽样值Q_k，即：

$$Q_k=\sum_{i=1}^{R}\sum_{j=1}^{W_i}Q_{ij}/SW \qquad (k=1, 2, \cdots, g) \quad (3\text{-}2\text{-}79)$$

$$SW=\sum_{i=1}^{R}W_i$$

式中 Q_k——第k个复合抽样值；

Q_{ij}——第i个专家的第j次抽样值；

W_i——第i个专家的权系数。

如此反复进行复合抽样，如果$g=2000$，则可得到2000个复合抽样值。最后以频率统计法求出石油资源量估计值的分布函数，从而给出各种概率下石油资源量的估计值。

加权抽样法算得的结果与概率加权法算得的结果相比，往往会有明显的区别。两种算法的区别在于，概率加权法给出的石油资源量估计值的范围较宽，而加权抽样法给出的范围较窄。所以，加权抽样法具有更强的综合能力，更能代表多数专家的评价意见。

最后还要指出，特尔菲班长的综合方法绝非限于上面讲的两种方法，应当根据实际情况，选定最合适的综合方法。

四、算　　例

对于表3-2-15中11位专家对沉积盆地所作的评价意见，特尔菲班长首先要对第6～9号4位专家的点估计值进行小区间展开。如果令（3-2-75）式中的$C=20$，则以C除以各专家的点估计值得到Δq值，例如其中8号专家的点估计值是70×10^8t，则有：

$$\Delta q=\frac{70}{20}=3.5\ (\times10^8\mathrm{t})$$

所以，第8号专家估计值展开后的端点为：

$$Q_{8min}=70-3.5=66.5\ (\times10^8\mathrm{t})$$

$$Q_{8max}=70+3.5=73.5\ (\times10^8\mathrm{t})$$

同样，对第6、7、9号专家点估计值也要分别展开为区间估计值。

1．按概率加权法计算

按概率加权法进行计算时，$SW=11$。而11位专家估计值中，最大及最小的估计值为：

$$Q_{max}=73.5\ (\times10^8\mathrm{t})$$

$$Q_{min}=20.15\ (\times10^8\mathrm{t})$$

按（3-2-76）、（3-2-77）、（3-2-78）式计算，该沉积盆地不同概率下石油资源量的估计值见表3-2-16，其分布函数见图3-2-29。

表3-2-16　概率加权法综合的石油资源量估计值表

概率（%）	石油资源量（×10⁸t）	概率（%）	石油资源量（×10⁸t）	概率（%）	石油资源量（×10⁸t）
100	20.1500	65	33.0812	30	48.4494
95	23.6947	60	34.0136	25	52.091[illegible]
90	26.0960	55	36.7326	20	57.417[illegible]
85	27.5879	50	40.1502	15	64.1822
80	29.0668	45	43.4057	10	67.0704
75	31.0514	40	45.2621	5	69.6500
70	32.1487	35	46.8505	0	73.5000

2．按加权抽样法计算

用加权抽样法计算的过程见图3-2-30。

加权抽样法的计算结果见表3-2-17，其分布函数见图3-2-31。

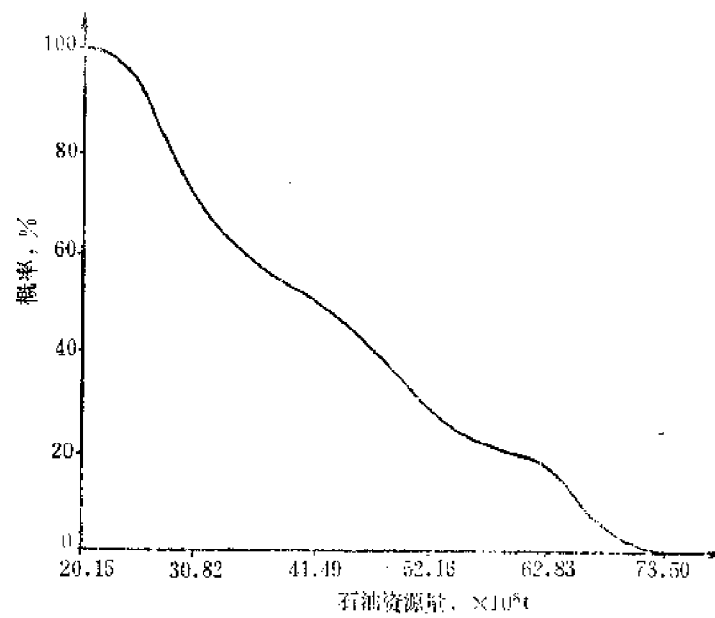


图3-2-29　概率加权法综合的石油资源量分布函数

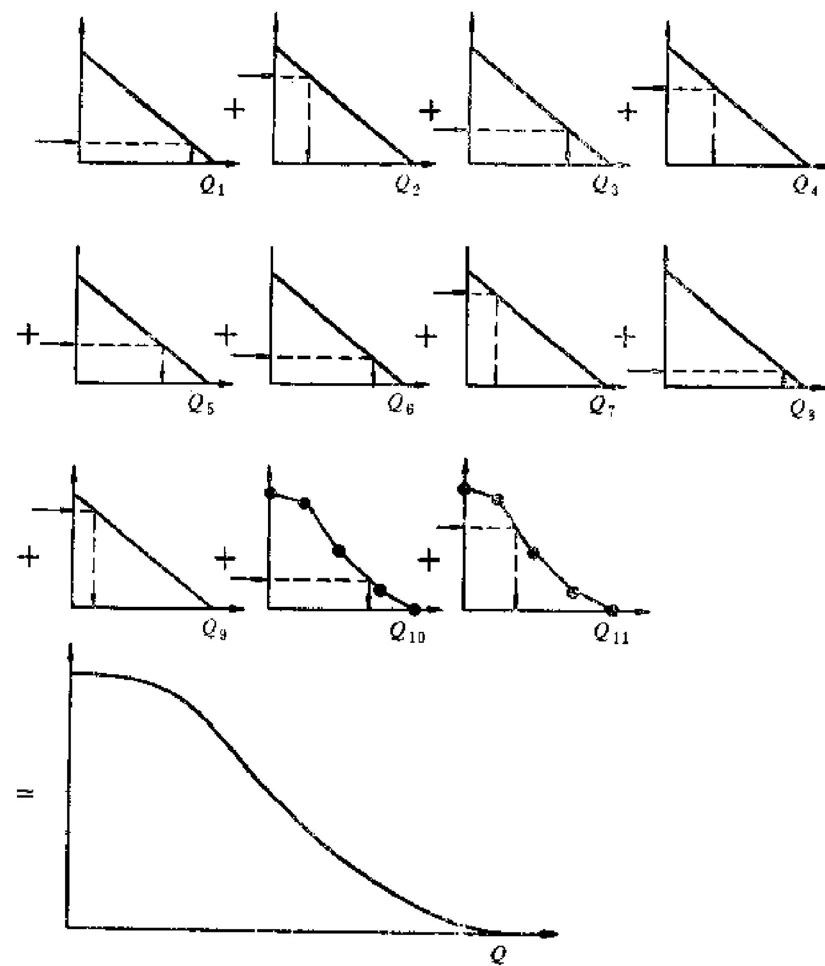


图3-2-30　加权抽样法计算过程的示意图

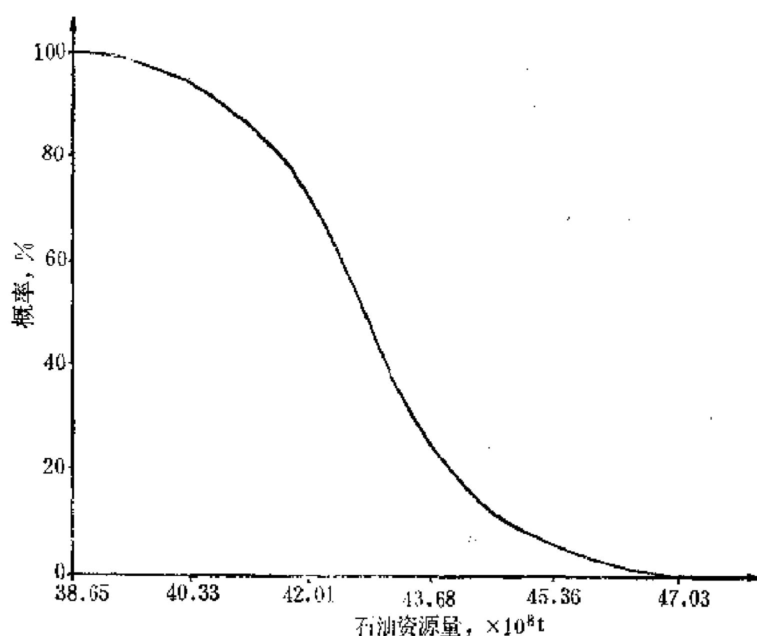


图3-2-31　加权抽样法综合的石油资源量分布函数

表3-2-17　加权抽样法综合的石油资源量估计值表

概率 (%)	石油资源量 ($\times 10^8$t)	概率 (%)	石油资源量 ($\times 10^8$t)	概率 (%)	石油资源量 ($\times 10^8$t)
100	38.6535	65	42.3428	30	43.5027
95	40.6200	60	42.5183	25	43.7017
90	41.1269	55	42.6797	20	43.9087
85	41.4547	50	42.8449	15	44.1566
80	41.7402	45	43.0084	10	44.4679
75	41.9646	40	43.1646	5	44.9182
70	42.1699	35	43.3341	0	47.0326

第三章　含油气有利地带的预测方法

目前众多的石油资源评价方法中，绝大多数的方法都属于石油资源量的预测方法，而预测探区含油气有利地带的方法很少。这也是现阶段石油资源评价工作中的一个薄弱环节。研究含油气有利地带的预测方法，不仅有利于提高评价结论的可靠性，而且可以指导探区当前的勘探工作，例如探区下一步的探井井位就是急需解决的实际生产问题。

第一节　模糊集合综合评价法

地质学中有许多含义模糊的概念，例如某一地质单元的含油气远景，某个地质圈闭的含油性等，其概念都是不准确的。因为构成这些概念的研究对象是没有确定边界的模糊体系，而模糊体系若用严格的数学方法处理，则可能会一筹莫展或者导出不完全真实的结果。

所谓模糊体系是指一些复杂的实际问题，它们不可能得到准确和明确的解答，因而需要用描述和分析的方法，来适应那些不准确的知识接界，或者适应我们主观上对实际问题有关价值的判断或评价。

石油勘探阶段，特别是早期石油勘探阶段，经过地面地质调查或地球物理勘探后，发现了一批地质圈闭，此时勘探人员最关心的问题就是这批圈闭中哪些是含油的，哪些是可能含油的，哪些是不含油的，这就是通常所说的地质圈闭的含油性评价问题。

诚然，某个地质圈闭中有没有石油，有多少石油储量，原本是地质历史演化的结果，而我们在研究这一地质圈闭有多少石油储量时，该圈闭的石油储量在客观上早已确定，但是，限于目前勘探手段所构成的观测系统的技术水平，或者观测精度不够，而使我们所能得到的这批地质圈闭含油性的映象却是一个模糊体系。因此，勘探人员依据这一模糊映象，不可能准确或明确地回答地质圈闭的含油性问题。

按以往的常规研究方法，通常是由观测到的地质信息加上勘探人员的实际经验，对每个地质圈闭进行打分，用以描述和分析地质圈闭与控制油气形成的因素之间关系，以适应人们对地质圈闭含油性评价的主观要求，这就是通常所说的按相对好坏给出地质圈闭含油性的排队评价。

美国控制论专家查德（L.A.Zaden）于1965年首先提出“模糊集合”的概念，对模糊体系用数学方法进行描述，从而创立了一个崭新的数学分支，即模糊数学。模糊数学是研究和处理模糊体系规律性的理论和方法，它把普通集合论只取0或1两个值的特征函数，推广到〔0，1〕区间上取值的隶属函数；把绝对的属于或不属于的“非此即彼”；扩张为更加灵活的渐变关系，因而便于把“亦此亦彼”中介过渡的模糊概念用数学方法处理。尽管目前模糊数学还不完善，在数学界也没有得到普遍承认，但是，模糊数学的思想方法与地质圈闭含油性的评价思路却十分相近。这就是以模糊数学方法对地质圈闭含油性进行综合评价的出发点。

应用模糊数学方法对地质圈闭含油性进行综合评价时，应考虑到如下四个问题：

（1）与地质圈闭含油性有关的多个控制油气形成的地质因素之间，可能存在多层次的

结构关系。例如，地质圈闭的含油性通常决定于生油条件、储油条件、盖层条件等，而这些基本地质条件，又由在当时勘探程度下可能取得的若干个次一级地质因素构成，例如生油条件可能与地质圈闭所处的生油条件分区、生油岩厚度、生油岩的地球化学指标等地质因素有关。

（2）对地质圈闭含油性进行综合评价时，对各个地质因素所起的作用很难给出确切的估计值，一般只能凭主观经验确定其相对重要程度，通常可用权重分配表示。

（3）设$X=\{x\}$是给定的论域，若论域X中任何一个元素x都有一个$\mu(x)$与之对应，并且满足$0\leqslant\mu(x)\leqslant1$，则称$\mu(x)$为隶属函数。需要指出，对如何建立隶属函数，目前还是模糊数学尚未完全解决的理论问题，在实际应用中，由于建立令人信服的隶属函数十分困难，只好以经验公式或者评语级别代替。

（4）矩阵合成运算中，仅用取大取小算子将会丢失很多信息，而使评价结果过于单调，甚至难于鉴别地质圈闭含油性的优劣。

一、地质圈闭的综合评价方法

地质圈闭的含油性是由多种地质条件所决定的。例如，某个探区经过地震勘探已经发现一批地质圈闭，钻探前需要进行圈闭排队，选择含油性最好的地质圈闭首先进行钻探。

如果评价地质圈闭含油性时，引用了n项地质因素，则可构成因素集合U

$$U=\{U_1,U_2,\cdots,U_i,\cdots,U_n\}$$

式中的$U_i(i=1,2,\cdots,n)$是集合U的元素或子集。当U_i是子集时，它可由n_i个元素或次一级子集组成，即

$$U_i=\{U_{i1},U_{i2},\cdots,U_{ij},\cdots,U_{in_i}\} \qquad (j=1,2,\cdots,n_i)$$

如果评价地质圈闭含油性时，预想分为m个级别，则可设评价集合V，即

$$V=\{V_1,V_2,\cdots,V_m\}$$

考虑到所引用的每项地质因素在评价地质圈闭含油性时所起的作用不同，可设A为因素集合U的权重分配，即

$$A=\{A_1,A_2,\cdots,A_i,\cdots,A_n\}$$

式中A_i是A的元素或子集。当A_i是子集时它可由n_i个元素或次一级子集组成

$$A_i=\{A_{i1},A_{i2},\cdots,A_{ij},\cdots,A_{in_i}\} \qquad (j=1,2,\cdots,n_i)$$

这里要求：

$$\sum_{i=1}^{n}A_i=1$$

$$\sum_{j=1}^{n_i}A_{ij}=1$$

从U到V的一个模糊映射$R(U_i)$叫作单项因素评价，$U_i\in U$时有

$$R(U_i)=(r_{i1},r_{i2},\cdots,r_{im})$$

对地质圈闭的含油性进行综合评价时，所引用的地质因素有时有准确的定量数据，有时只有相对关系或者定性描述。为了统一起见，在此一律采用相对评语表示子集$R(U_i)$。对于定量数据，可以通过等级变换转化为相对评语。

如评价集合分为好、中等、差3个级别时，可按表3-3-1中的评语级别表示子集$R(U_i)$。仿此，如评价集合分为好、较好、中等、较差、差5个级别时，可按表3-3-2中的评语级别表示子集$R(U_i)$。如若评价集合分为最好、好、较好、中等、较差、差、最差7个级别时，可按表3-3-3中的评语级别表示子集$R(U_i)$。

表3-3-1 3个级别的评语表

评语＼级别	−1	0	1
好	0	0.2	0.8
中等	0.25	0.5	0.25
差	0.8	0.2	0

表3-3-2 5个级别的评语表

评语＼级别	−2	−1	0	1	2
好	0	0	0	0.2	0.8
较好	0	0	0.2	0.6	0.2
中等	0	0.25	0.5	0.25	0
较差	0.2	0.6	0.2	0	0
差	0.8	0.2	0	0	0

表3-3-3 7个级别的评语表

评语＼级别	−3	−2	−1	0	1	2	3
最好	0	0	0	0	0	0.2	0.8
好	0	0	0	0	0.15	0.7	0.15
较好	0	0	0	0.2	0.6	0.2	0
中等	0	0	0.25	0.5	0.25	0	0
较差	0	0.2	0.6	0.2	0	0	0
差	0.15	0.7	0.15	0	0	0	0
最差	0.8	0.2	0	0	0	0	0

如果U_i是U的元素时，则可由n个模糊映射$R(U_i)$组成综合评价变换矩阵R

$$R=(r_{ij})_{n\times m}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix}$$

如果U_i是U的子集时，可由n_i个模糊映射

$$R(U_{ij})=(r_{ij1},r_{ij2},\cdots,r_{ijn_i})$$

组成单项地质因素的综合评价变换矩阵R_i

$$R_i=[r_{ijk}]_{n_i\times m}=\begin{pmatrix} r_{i11} & r_{i12} & \cdots & r_{i1m} \\ r_{i21} & r_{i22} & \cdots & r_{i2m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{in_i1} & r_{in_i2} & \cdots & r_{in_im} \end{pmatrix}$$

若U_i是U的元素，地质因素的权重分配A与综合评价变换矩阵R_h的合成B_h称为第h个地质圈闭的综合评价，即

$$B_h=A\bigcirc R_h \qquad (h=1,2,\cdots,p) \qquad (3\text{-}3\text{-}1)$$

若U_i是U的子集，而U_{ij}是U_i的元素时，首先要计算次一级地质因素的综合评价B_{hi}，即

$$B_{hi}=A_i\bigcirc R_{hi} \qquad (h=1,2,\cdots,p)\ (i=1,2,\cdots,n) \qquad (3\text{-}3\text{-}2)$$

（3-3-2）式中，B_{hi}的脚码i代表地质因素的编号，h代表地质圈闭的编号；○为合成算子。

次一级地质因素综合评价的计算结果要作为上一级综合评价变换矩阵的一行。

最后可用下面的（3-3-3）式求得每个地质圈闭含油性的综合评价值D_h，即

$$D_h=B_hC^T \qquad (h=1,2,\cdots,p) \qquad (3\text{-}3\text{-}3)$$

（3-3-3）式中的C^T是等级矩阵的转置矩阵。

当评价集合分为好、中等、差3个级别时，可令$C=[-1\ 0\ 1]$；当评价集合分为好、较好、中等、较差、差5个级别时，可令$C=[-2\ -1\ 0\ 1\ 2]$；当评价集合分为最好、好、较好、中等、较差、差、最差7个级别时，可令$C=[-3\ -2\ -1\ 0\ 1\ 2\ 3]$。

求出P个地质圈闭含油性的综合评价D_h后，则可按D值的大小进行地质圈闭含油性的排队，最后得到P个地质圈闭含油性相对好坏的次序。对于一个勘探地区，可以按计算得到的排队次序作为地质圈闭钻探的先后顺序。

二、矩阵合成时的运算规则

普通集合乘积矩阵的第i行第j列的元素值，等于左侧矩阵第i行元素与右侧矩阵第j列元素对应项乘积的代数和。但是，模糊矩阵合成的算子较多，除包括普通集合的乘法算子外，还可以根据实际需要定义多种其他算子。

1. 基本算子

假设a、b为模糊集合中的两个元素，这里定义了如下4种基本算子：

$$(1)\ a\vee b=\max(a,b) \qquad (3\text{-}3\text{-}4)$$

式中的$\max(a,b)$表示从a、b两个元素中选择数值大的元素值作为$a\vee b$的运算结果。

$$(2)\ a\vee b=\min(a,b) \qquad (3\text{-}3\text{-}5)$$

式中的$\min(a,b)$表示从a、b两个元素中，选择数值小的元素值作为$a\wedge b$的运算结果。

$$(3)\ a\cdot b=ab \qquad (3\text{-}3\text{-}6)$$

式中的ab表示a与b两个元素的乘积值，运算规则与普通集合的乘法运算一致。

$$(4)\ a\oplus b=\min(1,a+b) \qquad (3\text{-}3\text{-}7)$$

式中的$\min(1,a+b)$表示从1与元素和（$a+b$）中，选择数值小的元素作为$a\oplus b$的运算结果。

2. 矩阵合成时的运算方法

地质因素的权重分配A与综合评价变换矩阵R的合成称为地质圈闭的综合评价B，即

$$A \bigcirc R = B = (b_1, b_2, \cdots, b_m)$$

在矩阵合成运算时，可按实际需要选用合适的算子搭配方法。常用的算子搭配方法有如下4种：

（1）取小取大算法　这种算法可简记为（∧，∨），合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigvee_{i=1}^{n} (a_i \wedge r_{ij}) \quad (j=1,2,\cdots,m) \qquad (3\text{-}3\text{-}8)$$

（2）乘积取大算法　这种算法可简记为（·，∨），合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigvee_{i=1}^{n} (a_i \cdot r_{ij}) \quad (j=1,2,\cdots,m) \qquad (3\text{-}3\text{-}9)$$

(3)取小求和算法　这种算法可简记为（∧，⊕），合成矩阵B中的元素b_j的计算方法如下：

$$b_j = \bigoplus_{i=1}^{n} (a_i \wedge r_{ij}) \quad (j=1,2,\cdots,m) \qquad (3\text{-}3\text{-}10)$$

（4）乘积求和算法　这种算法可简记为（·，⊕），合成矩阵中的元素b_j的计算方法如下：

$$b_j = \bigoplus_{i=1}^{n} (a_i \cdot r_{ij}) \quad (j=1,2,\cdots,m) \qquad (3\text{-}3\text{-}11)$$

例如：$A = (0.6 \quad 0.3 \quad 0.2)$

$$R = \begin{pmatrix} 0.2 & 0.3 & 0.5 \\ 0.4 & 0.2 & 0.4 \\ 0.1 & 0.6 & 0.3 \end{pmatrix}$$

按（∨，∨）计算时，$B = (0.3 \quad 0.3 \quad 0.5)$

按（·，∨）计算时，$B = (0.12 \quad 0.18 \quad 0.3)$

按（∧，⊕）计算时，$B = (0.6 \quad 0.7 \quad 1.0)$

按（·，⊕）计算时，$B = (0.26 \quad 0.36 \quad 0.48)$

三、算　例

〔1〕根据地质调查及地震勘探，在某探区内发现了5个地质圈闭，见表3-3-4。

从表3-3-4中可看出，由于该探区的勘探程度较低，所以生油条件、储油条件、盖层条件都是由定性的评语描述的；构造条件是用地震资料定量描述的。

根据勘探人员的认真分析，各地质因素的权重分配如下：

综合评价
- 生油条件　(0.25)
- 储油条件　(0.25)
- 盖层条件　(0.15)
- 构造条件 (0.35)
  - 面　积　(0.4)
  - 幅　度　(0.3)
  - 断层情况　(0.3)

表3-3-4　地质圈闭的各地质因素评语及数据表

地质因素 \ 圈闭编号		1	2	3	4	5
生油条件		较差	中等	较好	较好	中等
储油条件		中等	较好	中等	中等	较差
盖层条件		较好	中等	中等	较好	好
构造条件	面积(km^2)	10	20	10	15	20
	幅度（m）	100	200	100	150	90
	断层情况	2条	1条	无	无	1条

构造条件中的面积、幅度、断层情况是定量数据，为了能与生油条件、储油条件、盖层条件的定性评语搭配使用，可按表3-3-5中规定的标准分为5个等级。该表中的标准是由熟悉探区地质情况的勘探人员讨论商定的。需要指出，每个探区必须根据具体的情况制定分级标准。

表3-3-5　各构造因素的评语分级标准

地质因素 \ 评语	差	较差	中等	较好	好
构造面积(km^2)	<5	5～10	10～30	30～50	>50
构造幅度(m)	<50	50～100	100～200	200～300	>300
断层情况	>2条	1～2条	1条	无	无

按表3-3-5规定的标准，表3-3-4中的定量数据都可以转换成相应的评语，转换后的评语见表3-3-6。

表3-3-6　地质圈闭的各地质因素评语表

地质因素 \ 圈闭编号		1	2	3	4	5
生油条件		较　差	中　等	较　好	较　好	中　等
储油条件		中　等	较　好	中　等	中　等	较　差
盖层条件		较　好	中　等	中　等	较　好	好
构造条件	面积	较　差	中　等	较　差	中　等	中　等
	幅度	较　差	中　等	较　差	中　等	较　差
	断层情况	较　差	中　等	好	好	中　等

这里选用的地质因素是有层次关系的，生油条件U_1、储油条件U_2、盖层条件U_3、构造条件U_4，总共4项地质因素构成因素集合U

$$U=(U_1,U_2,U_3,U_4)$$

其中U_1、U_2、U_3是因素集合中的元素，而构造条件U_4是子集。子集U_4是由构造面积U_{41}、构造幅度U_{42}、断层情况U_{43}这3项次一级的地质因素组成，即

$$U_4=(U_{41},U_{42},U_{43})$$

由于各项地质因素及次一级构造因素评语都是按5个级别划分的，所以可构成5个级别的评语集合V。

$$V=(V_1,V_2,V_3,V_4,V_5)$$

因为地质因素分为两级，所以也有如下两级相应的权重分配。

$A=(0.25\quad 0.25\quad 0.15\quad 0.35)$

$A_4=(0.4\quad 0.3\quad 0.3)$

为了对这5个地质圈闭进行排队，首先要从次一级的构造因素，即构造面积、构造幅度、断层情况开始计算。

R_{14}是第1个圈闭的第4项地质因素（构造条件）的综合评价变换矩阵，按（·,⊕）乘积求和算法计算时，第一个圈闭的第4项地质因素的综合评价B_{14}为：

$$B_{14}=A_4\bigcirc R_{14}=(0.4\quad 0.3\quad 0.3)\bigcirc\begin{pmatrix}0.2 & 0.6 & 0.2 & 0 & 0\\0.2 & 0.6 & 0.2 & 0 & 0\\0.2 & 0.6 & 0.2 & 0 & 0\end{pmatrix}$$

$$=(0.2\quad 0.6\quad 0.2\quad 0\quad 0)$$

按同样的方法计算，可以得到第2、3、4、5号地质圈闭的构造条件综合评价：

$B_{24}=(0\quad 0.25\quad 0.5\quad 0.25\quad 0)$

$B_{34}=(0.14\quad 0.42\quad 0.14\quad 0.06\quad 0.24)$

$B_{44}=(0\quad 0.175\quad 0.35\quad 0.235\quad 0.24)$

$B_{54}=(0.06\quad 0.355\quad 0.41\quad 0.175\quad 0)$

继续作地质圈闭的综合评价计算时，上面算得的R_{h4}要作为综合评价变换矩阵R_h中的第4行。如果仍以（·，⊕）乘积求和算法计算，则第1个地质圈闭的综合评价B_1为：

$$B_1=A\bigcirc R_1$$

$$=(0.25\quad 0.25\quad 0.15\quad 0.35)\bigcirc\begin{pmatrix}0.2 & 0.6 & 0.2 & 0 & 0\\0 & 0.25 & 0.5 & 0.25 & 0\\0 & 0 & 0.2 & 0.6 & 0.2\\0.2 & 0.6 & 0.2 & 0 & 0\end{pmatrix}$$

$$=(0.12\quad 0.4225\quad 0.275\quad 0.1525\quad 0.03)$$

按同样方法计算，可以得到第2、3、4、5号地质圈闭的综合评价：

$B_2=(0\quad 0.1857\quad 0.425\quad 0.3375\quad 0.07)$

$B_3=(0.049\quad 0.247\quad 0.299\quad 0.271\quad 0.134)$

$B_4=(0\quad 0.12375\quad 0.3275\quad 0.38475\quad 0.164)$

$B_5=(0.071\quad 0.33675\quad 0.3185\quad 0.15375\quad 0.12)$

最后，按（3-3-3）式计算每个地质圈闭的综合评价值$D_h(h=1,2,3,4,5)$，其中D_1为

$$D_1=(0.12\quad 0.4225\quad 0.275\quad 0.1525\quad 0.03)\begin{pmatrix}-2\\-1\\0\\1\\2\end{pmatrix}=-0.45$$

按同样方法计算，可以得到第2、3、4、5号地质圈闭的综合评价值：

$D_2=0.29$

$D_3=0.194$

$D_4=0.589$

$D_5=-0.085$

按以上计算结果，5个地质圈闭的排队次序应为：D_4，D_2，D_3，D_5，D_1，其中第4号地质圈闭的含油气地质条件最好，第1号地质圈闭的含油气地质条件最差。

〔2〕二连盆地中马尼特坳陷至1984年共发现119个局部构造，按地震反射层T_5、T_7、T_{11}（浅、中、深）收集了如下地质参数：

（1）构造的圈闭面积（km^2）；

（2）构造的圈闭幅度（m）；

（3）构造的类型，数字化后，背斜赋值为3，半背及断背斜为2，断块及其他为1；

（4）生油相带是构造圈闭所处的生油凹陷位置，数字化后，Ⅰ类生油区内的构造圈闭赋值为3，Ⅰ至Ⅱ类生油区之间的为2，其他区内的为1。

119个局部构造先按T_5、T_7、T_{11}三个单层计算后，再对局部构造进行最后的综合评价。表3-3-7列出了119个局部构造中的前20名含油气地质条件较好的局部构造排队顺序及综合评价值。

表3-3-7　二连盆地马尼特坳陷局部构造的综合评价结果表

排队序号	构造名称	综合评价值	排队顺序	构造名称	综合评价值
1	阿尔善	2.040	11	哈　邦	−0.960
2	哈达图	0.560	12	邦　东	−0.960
3	蒙古林	0.160	13	萨音乌苏	−0.960
4	贡　尼	0.000	14	巴　润	−1.120
5	哈　南	−0.200	15	巴　东	−1.120
6	准沟东	−0.760	16	毛　普	−1.120
7	额尔熟切	−0.760	17	吉　北	−1.120
8	哈　北	−0.800	18	莎东1	−1.120
9	准　沟	−0.960	19	莎东2	−1.120
10	沃布多东	−0.960	20	毛普南	−1.320

第二节　多种信息叠合评价法

多种信息叠合评价法是对已掌握的探区地质资料进行综合处理的一种方法。通过叠合处理可以得到与含油气有利地带关系密切的综合地质信息，因而有利于制定探区的勘探方案。

含油气有利地带的预测，不仅仅限于钻探井位的选择，而且应包括全国范围内的有利含油气盆地的选定；沉积盆地内有利地质坳陷的挑选；地质坳陷内有利凹陷或凸起的确定；凹陷或凸起上有利地质圈闭的排队以及地质圈闭上最佳钻探井位的圈定等一系列预测工作。

按传统的石油地质研究方法，为确定含油气有利地带，首先要分别研究生油、储油、盖层、运移、聚集、保存等控制油气形成的基本地质条件，最后再通过综合研究工作进行探区内含

油气有利地带的预测。

一、多种信息叠合评价法的要点

一般情况下，一个新的探区总会有或多或少的地质资料。这些不同类型的地质资料（包括地质数据、地质图件、地质观点），都应当看作是从不同侧面向地质人员提供的寻找有利勘探地带的地质信息。

多种信息叠合评价法的基本思路，是把控制油气形成的各种不同的单一地质因素看作是基础地质信息，表示基础地质信息的地质图件称为基础地质信息图件；由若干个基础地质信息图件经叠合生成的图件称为组合地质信息图件；再由若干组合地质信息图件经二次叠合生成的图件称为综合地质信息图件。

可见，多种信息叠合评价法的要点可以概括为"图加图出新图"。新图中的信息是由基础地质信息经过逐级多次叠合生成的复合信息。

二、多种信息叠合评价法的实施步骤

1. 地质数据的归类与分级

地质数据的归类与分级是将已收集到的地质数据，形成归类合理的、层次分明的数据结构。一般情况下，可分为两级，即首先由同类的基础地质信息构成组合地质信息，再由组合地质信息构成综合地质信息。

例如在某个探区已收集到生油岩厚度、生油岩有机碳含量、储集层厚度、储集层孔隙度、储集层渗透率、盖层厚度、局部构造特征等总共7项与油气形成有关的地质数据，它们都是基础地质信息。其中的生油岩厚度与生油岩有机碳两项基础地质信息同属生油条件，因此生油条件就是由这两项基础地质信息构成的组合地质信息。同样，储集层厚度、孔隙度、渗透率三项基础地质信息构成了储油条件这个组合地质信息。归类后可以得到有层次关系的数据结构，见表3-3-8。

表3-3-8 地质信息结构表

组合地质信息	基础地质信息
生油条件	生油岩厚度、生油岩有机碳含量
储油条件	储集层厚度、孔隙度、渗透率
盖层条件	盖层厚度
构造条件	局部构造特征（例如构造面积、闭合度等）

2. 生成基础地质信息图件

如果某些基础地质信息已有现成的图件并且符合要求，则不必由基础地质数据去生成图件。但是，在多数情况下，由于图件比例尺不同，以及图件的收缩、变形、破损等原因，往往需要重新生成图件，即便是直接将原图输入计算机，也要作位置校正及误差校正，否则叠合后的地质信息将会失真。

由基础地质数据生成平面图件就是用约定的插值计算方法，由计算机绘制等值线图或分带图。具体的实现方法很多，例如距离倒数平方加权法就是常用的平面插值方法。

3. 生成组合地质信息叠合图件

由同类的基础地质信息图件，按约定的算法可以生成组合地质信息图件。例如由生油岩有机碳含量与生油岩厚度的等值线图或分带图，叠合后可以生成生油条件等值线图或分带图。

在叠合前，可以按每种基础地质信息在组合地质信息中所起作用的大小，赋以相应的权系数，使各个基础地质信息起到相应的作用，从而使生成的组合地质信息图更为合理。

4. 生成综合地质信息的二次叠合图

由若干个组合地质信息图件，按约定的叠合方法可以生成最后的综合地质信息图件。由基础地质信息图件到形成综合地质信息图件，总共经过了两次叠合过程。

同样，实行二次叠合前，仍然需要根据每个组合地质信息对探区油气形成所起作用的大小，赋以适当的权系数，以保证每种组合地质信息在最终评价上起到应有的作用。

三、地质数据的平面插值

在地质数据中，除地球物理、遥感等少数地质数据是密集采样外，其他绝大多数的地质数据都是稀疏的离散点值。因此，为了进行地质信息间的叠合，事先需要根据少数的离散点值扩充为平面密集的数值点。这一问题在数学上就是二维（平面）数据的插值问题。

这里介绍一种简单而适用的平面插值方法，即距离倒数平方加权法。如果探区中有m种可作为指导找油的基础地质数据，其中第j种基础地质数据经过挑选后可供使用的原始数据有n个，其第i个数据的平面坐标为（x_i，y_i），观测值为z_i。设平面插值域上任意一点p的平面坐标为（x,y），p点与第i个原始数据点的距离为D_i，则

$$D_i=[(x-x_i)^2+(y-y_i)^2]^{\frac{1}{2}} \qquad (i=1,2,\cdots,n) \qquad (3\text{-}3\text{-}12)$$

p点的插值z可用下面的公式求得

$$z=\begin{cases}\sum\limits_{i=1}^{n} z_i(D_i)^{-2}\Big/\sum\limits_{i=1}^{n}(D_i)^{-2} & （所有的D_i\neq 0时）\\ z_i（如果有D_i=0时）\end{cases} \qquad (3\text{-}3\text{-}13)$$

由（3-3-13）式可以看出，平面上任意一点p的插值方式有两种可能：当p点与第i个原始数据点重合，即如果$D_i=0$时，则令z值等于z_i；当p点与n个原始数据点都不重合，即如果所有的$D_i\neq 0$时，则插值z受n个原始数据z_i的影响，而每个原始数据对z的影响程度，与原始数据点到p点的距离平方成反比。可见，z值的大小主要受靠近p点的原始数据影响，而远离p点的原始数据对z的影响较小。

需要说明的是，上面的插值法仅仅是最简单的一种方法。根据实际需要，计算时也可以约定其他插值方法，例如高次趋势面逼近法、克里格法、埃米尔插值法等。

由于各个探区的地质情况千差万别，即使是同一个探区中，也有若干个在性质上不同的次一级地质单元，因此，为了使插值结果符合实际地质情况，在插值时要对原始数据点的使用数量作一些人为限定，限定的方式大体有如下几种：

1. 全点插值法

全点插值法是对原始数据不加任何限定，即所有数据点都参加插值计算。当原始数据点比较少，而勘探人员对探区的地质结构又不太清楚时，最好用全点插值法。用这种插值法得到的图形比较光滑，不容易发生图形畸变。

2. **近点插值法**

近点插值法是从n个原始数据点中选用距离p点最近的t个点参加插值计算。当探区的地质结构比较复杂时，为了使参加插值计算的数据点限定在同一地质单元内，最好使用近点插值法。近点的个数t值，可视具体的地质情况确定。

使用近点插值法时，必须首先计算插值点到所有原始数据点之间的距离，并且要按距离的远近，从小到大依次排序，而后从中选出t个与p点距离最近的原始数据点。这个排序计算要占用一定的计算时间。

3. **圆内插值法**

圆内插值法是以插值点p为圆心，选择合适的半径R划圆，以圆周为界，在圆周以内的点叫圆周内点，插值时只使用圆周内点的原始数据参加计算，见图3-3-1。

在图3-3-1中，以p点为圆心以R为半径的圆内共有4个原始数据点，它们参加插值计算。

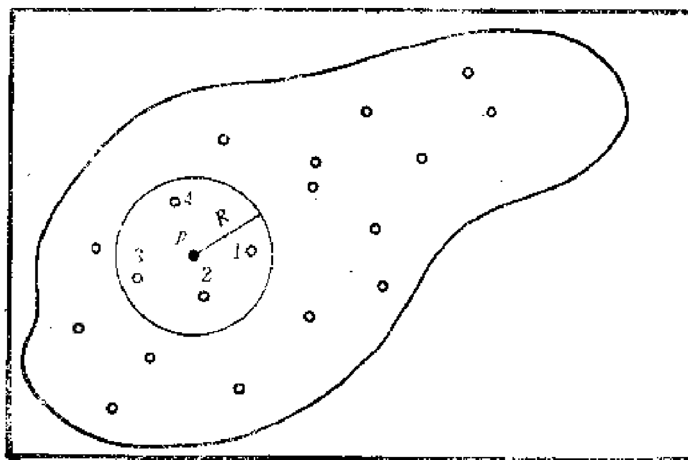


图3-3-1 圆内插值法示意图

4. **象限插值法**

象限插值是以插值点p为中心，将插值域分为若干个象限，一般可分为4个象限或8个象限，在每个象限中选择距p点最近的t个原始数据点参加计算。这种插值方法可以保证所使用的数据点在各象限方向上的均匀性，以消除由于数据点分布不均匀造成的方向性干扰，见图3-3-2。

需要指出，对原始数据点的使用数量以及选点方法都是人为约定的，所以绝非仅有以上4种方法，而应当按实际需要灵活约定。

在以上4种插值方法中，相比之下，圆内插值法相对好些，它有以下两条优点：

（1）圆内插值法能较好地体现勘探程度，如果插值点附近（以R为半径的圆周内）原始数据点较多时，插值结果的可靠性要高些；而原始数据点较少时，插值结果的可靠性则低些，当插值圆内无原始数据时，因无法插值而自然形成插值的空白区。

（2）由于圆内插值法既避开了全点插值法中要用全部原始数据点进行插值计算，又避开了近点插值法中要计算插值点到全部原始数据点的距离以及按距离大小的排序过程，也避开了象限插值法中的象限划分以及象限内选近点的计算过程，因此，圆内插值法的计算时间相

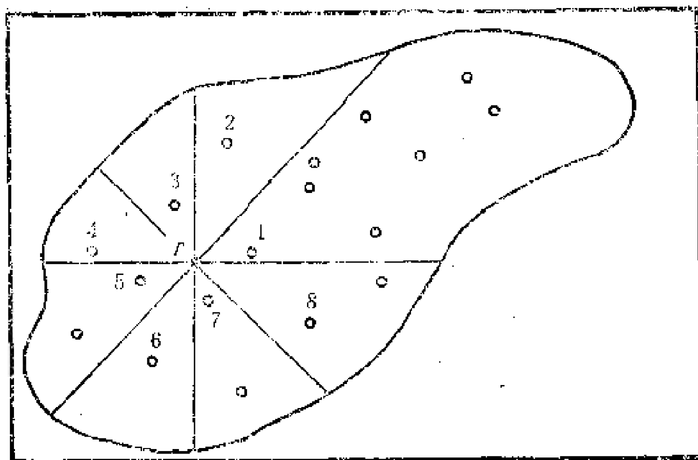


图3-3-2　象限插值法示意图

对较短。

四、多种信息的叠合方法

各种基础地质数据经过插值计算后，就可以得到等值线图或分带图。如果按已归类的数据结构，把同一类的基础地质信息图件重叠在一起，按某种约定的算法进行叠合，则可以得到组合地质信息图。

再经过二次叠合就能得到最终的预测探区含油气有利地带的多种信息叠合评价图。

1．叠合前的数据预处理

（1）地质数据的正规化　进行信息叠合前，每种基础地质数据都要经过标准化处理，使它们变换到同一尺度范围内，以保证各种地质信息之间的等价性。这里建议使用极差正规化方法，将各种原始数据变换到〔0，1〕区间范围内。

（2）对地质信息赋权值　进行信息叠合之前，勘探人员应当根据每种基础地质信息或组合地质信息对油气形成所起的作用，赋以合理的权系数，以体现叠合时它们所起的作用。

2．一次叠合方法

一次叠合方法是指由基础地质信息叠合生成组合地质信息的方法。

（1）乘积叠合　这种叠合方法是把平面上同一坐标点的m种基础地质信息值进行连乘，得到组合地质信息z_j，即

$$z_j=\prod_{i=1}^{m} z_{ij} \qquad (j=1,2,\cdots,w) \tag{3-3-14}$$

式中　z_j——经过一次叠合生成的第j种组合地质信息；

z_{ij}——第j种组合地质信息中的第i种基础地质信息。

用乘积叠合图对探区进行分带评价时，其分带区间是不等间距的，区间间隔值可按如下公式计算

$$H_r=\left[\frac{1}{k}(r-1)\right]^m \qquad (r=1,2,\cdots,k+1) \tag{3-3-15}$$

式中　H_r——第r个分带区间的间隔值；

k——分带区间总数；

m——进行叠合的基础地质信息个数。

（2）累加叠合　这种叠合方法是把平面上同一坐标点的m种基础地质信息值进行累加，得到组合地质信息z_j，即

$$z_j=\sum_{i=1}^{m} z_{ji} \qquad (j=1,2,\cdots,w) \tag{3-3-16}$$

对于累加叠合，其分带区间是等间距的，区间间隔值的计算公式如下：

$$H_r=\left[\frac{1}{k}(r-1)\right]m \qquad (r=1,2,\cdots,k+1) \tag{3-3-17}$$

（3）集合取小叠合　这种叠合方法是把平面上同一坐标点的每种基础地质信息作为一个元素，如果有m种基础地质信息，则可构成一个由m个元素组成的一个集合。叠合时是从集合中取出数值最小的元素作为叠合值，即

$$z_j=\min_{1\leqslant i\leqslant m}(z_{ji}) \qquad (j=1,2,\cdots,w) \tag{3-3-18}$$

对于集合取小叠合，其分带区间也是等间距的，区间间隔值的计算公式如下：

$$H_r=\frac{1}{k}(r-1) \qquad (r=1,2,\cdots,k+1) \tag{3-3-19}$$

3. 二次叠合方法

二次叠合是指经过一次叠合生成的w种组合地质信息，再进行二次叠合，生成综合地质信息。如果归类后的地质信息有两级，那么，由基础地质信息形成最后的综合地质信息，按上述三种叠合方法的组合，总共有9种叠合方法。

（1）双重乘积叠合　这种叠合是指由基础地质信息经过乘积叠合生成组合地质信息，再由组合地质信息经过乘积叠合生成综合地质信息。其计算公式为：

$$z=\prod_{j=1}^{w} z_j=\prod_{j=1}^{w}\prod_{i=1}^{m} z_{ji} \tag{3-3-20}$$

式中　z——综合地质信息；

z_j——第j种组合地质信息；

z_{ji}——第j种组合地质信息中的第i种基础地质信息。

（2）乘积累加叠合

$$z=\prod_{j=1}^{w} z_j=\prod_{j=1}^{w}\sum_{i=1}^{m} z_{ji} \tag{3-3-21}$$

（3）乘积取小叠合

$$z=\prod_{j=1}^{w} z_j=\prod_{j=1}^{w}\left(\min_{1\leqslant i\leqslant m} z_{ji}\right) \tag{3-3-22}$$

（4）双重累加叠

$$z=\sum_{j=1}^{w} z_j=\sum_{j=1}^{w}\sum_{i=1}^{m} z_{ji} \tag{3-3-23}$$

（5）累加乘积叠合

$$z=\sum_{j=1}^{w} z_j=\sum_{j=1}^{w} \prod_{i=1}^{m} z_{ij} \tag{3-3-24}$$

（6）累加取小叠合

$$z=\sum_{j=1}^{w} z_j=\sum_{j=1}^{w} (\min_{1\leqslant i\leqslant m} z_{ij}) \tag{3-3-25}$$

（7）双重取小叠合

$$z=\min_{1\leqslant j\leqslant w} (\min_{1\leqslant i\leqslant m} z_{ij}) \tag{3-3-26}$$

（8）取小乘积叠合

$$z=\min_{1\leqslant j\leqslant w} \left(\prod_{i=1}^{m} z_{ij}\right) \tag{3-3-27}$$

（9）取小累加叠合

$$z=\min_{1\leqslant j\leqslant w} \left(\sum_{i=1}^{m} z_{ij}\right) \tag{3-3-28}$$

4. 各种叠合方法的地质含义

乘积、累加、集合取小叠合方法的地质含义是不相同的。

乘积叠合方法，适用于被叠合的各种地质信息的乘积值大体上可以代表一个新的地质变量或者具有某种特定的地质含义的情况。例如，生油岩的厚度与氯仿沥青含量的乘积值，是与生油潜量有关的组合地质信息，大体上相当于生油丰度。

累加叠合方法，适用于各种被叠合的地质信息之间的关系尚不清楚，而经累加叠合后的组合地质信息表示各种与油气形成有关的地质信息的总和。累加叠合后的数值越大说明含油气的可能性越大。

集合取小叠合方法是出于最小因素思想，例如在一个勘探地区，如果生油、储油、盖层等控制油气形成的必要地质条件中，只要其中有一项不具备油气形成条件，则形成不了油气藏。可见，集合取小叠合是从最保险的角度出发的。

诚然，使用哪种叠合方法进行一次叠合，使用哪种叠合方法进行二次叠合，以及在叠合前对基础地质信息或组合地质信息所赋的权系数的大小，都要根据具体地质条件，由熟悉探区情况的勘探人员讨论商定。

五、算　例

我国北方某一沉积盆地是一个中新生代迭合的断陷坳陷盆地，其沉积岩厚度达到5000m以上，具备形成油气的基本地质条件，有条件建成一个新的产油区。

盆地中的M坳陷是最有远景的含油地区，因而评价该坳陷内各个地带的含油气地质条件，以及在坳陷内寻找最有利的勘探地带是十分重要的。目前，勘探的主要目的层系是k系p组，钻探工作主要集中在坳陷的东部地区。但是，全坳陷已基本完成地震勘探工作，因此用多种信息叠合法进行评价时，是以地震勘探资料为主，再加上已有的钻井资料，对全坳陷进行了有利勘探地带的预测。

根据地震资料及钻井资料，我们选用生油岩厚度、TTI值、生油岩沉积相，储集层厚

度、储集层沉积相，局部构造面积、幅度、类型，钻井油气显示，盖层厚度等总共10项基础地质信息。由这10项基础地质信息经过一次累加叠合后生成生油条件、储油条件、构造条件、含油气状况、盖层条件总共5项与油气形成有关的组合地质信息。

根据地震、钻井资料，在M坳陷已发现了76个局部构造。所选用的10种基础地质信息，按统一标准划分为一、二、三总共3个级别，分级标准与地质信息间的数据结构见表3-3-9，表中的3个级别中三级的含油气条件最好，二级居中，一级最差。

表3-3-9 M坳陷的地质信息结构及分级标准表

组合地质信息	基础地质信息	一　级	二　级	三　级
生油条件	生油岩厚度（m）	<300	300～500	>500
	TTI值	<8	>256	8～256
	生油岩沉积相	其他相	浅湖相	深　湖
储油条件	储集层厚度（m）	<100	100～300	>300
	储集层沉积相	湖沼相	河流平原相	冲积扇三角洲相
构造条件	构造面积(km^2)	<20	20～35	>35
	构造幅度（m）	<100	100～150	>150
	构造类型	断块、岩性	断鼻、半背斜、潜山	背　斜
含油气状况	油气产状	干　井	油气显示	油气流
盖层条件	盖层厚度（m）	<1200	>2400	1200～2400

为了计算方便，将基础地质信息按一、二、三级分别赋值为1、2、3。按分级标准，将M坳陷76个局部构造的原始数据经过数值化后输入计算机。各种原始数据的坐标均以原图的左下角为坐标原点（0，0），各点的纵、横坐标值是以cm为单位在原图上实际测量的。

5种组合地质信息经二次累加叠合生成最终的多种信息叠合评价图。一次以及二次叠合过程见图3-3-3。

叠合计算时选用的计算参数见表3-3-10。按这些选定的参数，经过计算得到10张基础地质信息的平面插值图；由这10张图经过一次累加叠合生成5张组合地质信息图，即：生油条件评价图、储油条件评价图、构造条件评价图、含油气状况图、盖层条件评价图；由这5张图经过二次累加叠合生成的多种信息叠合评价图，见图3-3-4。为了便于了解76个局部构造与评价结果的对应关系，图3-3-5中给出了局部构造的号码及其所在位置。

这些评价图统一划分为5个级别的分带，即：含油气最有利地带、有利地带、中等地带、不利地带、最不利地带。

根据计算结果，选出了21个含油气地质条件较好的局部构造，它们的号码是4、15、16、28、30、31、32、33、34、41、42、43、49、62、63、68、69、70、71、74、76。其中的30、31、68、69号局部构造经钻探已证实为含油构造。5年后，图3-3-4中东部的最有利与有利地区，已建成具有一定规模油气产能的油田区。

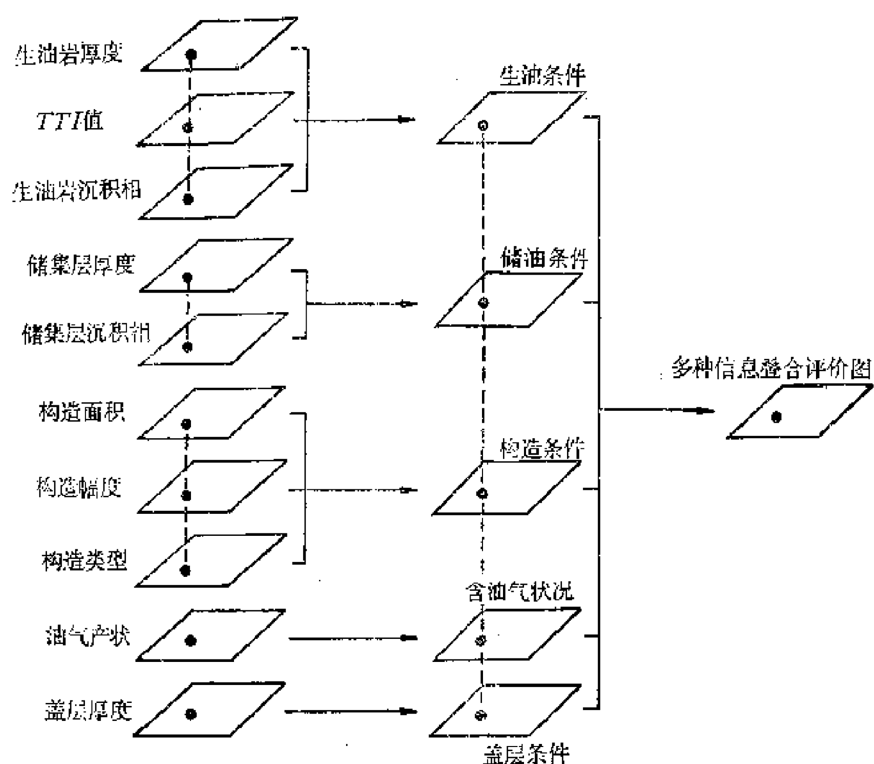


图3-3-3 地质信息叠合过程示意图

表3-3-10 叠合计算时的参数表

基础地质信息名称	基础地质信息个数	平面插值方法	一次叠合时的权系数	一次叠合方法	组合地质信息名称	组合地质信息中的基础地质信息个数	二次叠合时的权系数	二次叠合方法
生油岩厚度	76	圆内法	1.0	累加叠合	生油条件	5	1.0	累加叠合
TTI值	76	圆内法	1.0					
生油岩沉积相	76	圆内法	1.0					
储集层厚度	76	圆内法	1.0	累加叠合	储油条件	2	1.0	
储集层沉积相	76	圆内法	1.0					
构造面积	76	圆内法	1.0	累加叠合	构造条件	3	1.0	
构造幅度	76	圆内法	1.0					
构造类型	76	圆内法	1.0					
油气产状	76	圆内法	1.0	累加叠合	含油气状况	1	1.0	
盖层厚度	76	圆内法	1.0	累加叠合	盖层条件	1	1.0	

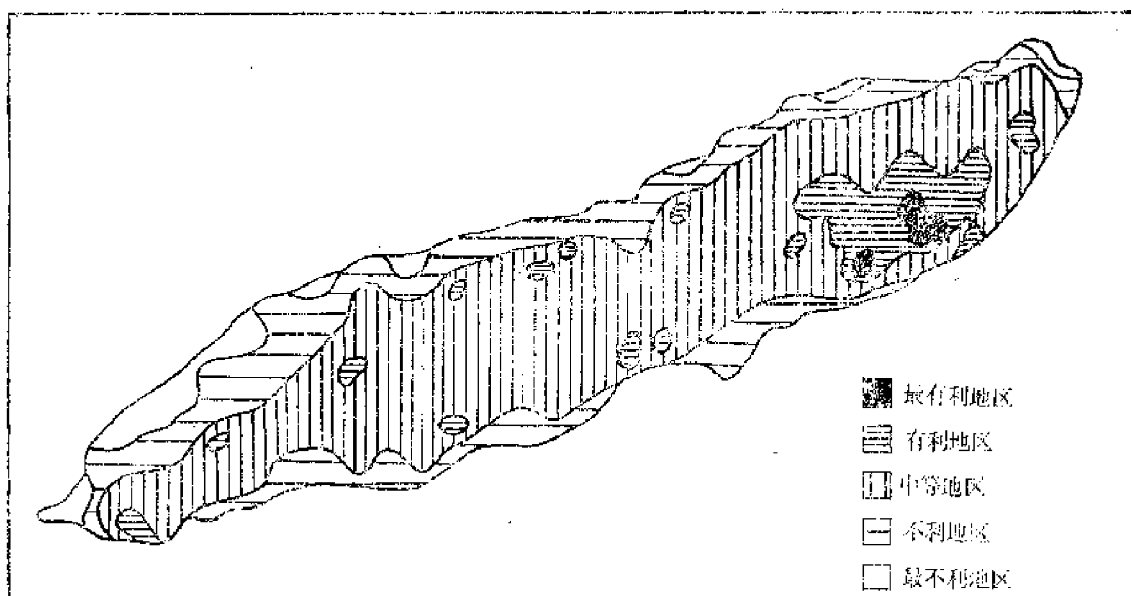


图3-3-4　多种信息叠合评价图

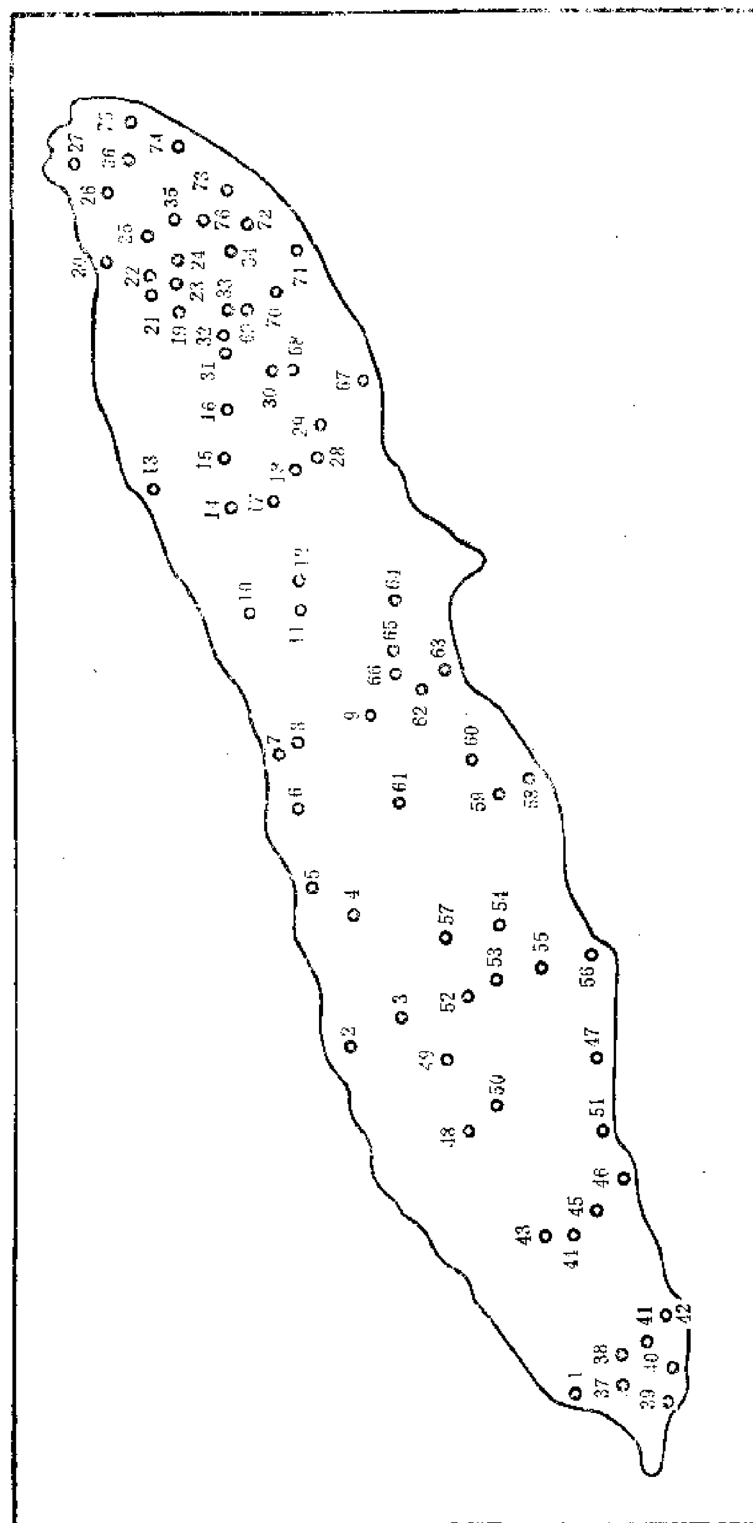


图3-3-5 局部构造位置图

第四章　经济评价与决策分析方法

石油勘探的经济评价是指从经济角度分析勘探结果的效益问题。而决策分析是指从若干个勘探方案中选择最佳勘探方案的问题。

第一节　勘探方案的线性规划模型

应用线性规划方法制定勘探方案，可以按总体设想从全局观点出发，把有关因素尽可能考虑进去，例如，探区中潜在的石油资源量、单位储量的综合勘探成本、方案实施期间的产量规划、探区中各分区之间的产量分配、不同类型储量之间的产量分配等因素都与勘探方案有关。这些因素就是线性规划模型中的约束条件，在此条件下求解目标函数就能得出总体设想的数值解，包括勘探周期、投资费用以及经济效益等。应用线性规划方法可以在短期制定出多种勘探方案，通过勘探方案对比可以选择出最佳勘探方案，以达到在最短时间内，用最少的投资，获得最大的经济效益。

石油勘探方案是指某一勘探地区（例如，一个地质坳陷、一个沉积盆地、一个省、一个国家等）、在某段时间内（例如，一年、一个五年计划期间等）勘探工作的总体部署及阶段安排，它包括勘探费用的投资分配，勘探设备的运行调拨，计划投入的勘探工作量及其具体部署等。因此，制定一个勘探方案将会涉及到多方面的因素，如对勘探地区的地质认识，计划新增产能及年度配产，已开发油田的产量递减，探区中不同分区的勘探成本等。

一、制定勘探方案的基本原则以及应当考虑的因素

一般情况下，应以投入最少的勘探工作量（或转成的货币值），在预定时间内完成或超额完成国家、上级主管部门下达的油气产量任务作为制定勘探方案的基本原则。

根据我国的目前实际情况，制定一个勘探方案时，需要考虑的因素大体有如下几个方面。

1．探区中潜在的油气资源量

探区中潜在的油气资源量是制定勘探方案的物质基础。潜在的油气资源量是指基于目前对探区的地质认识，按各种定量方法预测出的、并已得到有关人员认可的不同级别、不同类型的油气资源总量。

考虑到勘探成本上的差别，油气资源量可按勘探深度划分为若干种类型，如可划分为勘探深度小于2000m的油气资源，2000m到3500m的油气资源，以及大于3500m的油气资源等；也可以按预测的油气田规模划分为大于100Mt的油气资源，100Mt至20Mt的油气资源，小于20Mt的油气资源等。对于这些不同类型的油气资源，不仅其勘探成本不一样，而且勘探方法也不一样。

特别需要指出，对于探区的潜在资源量必须考虑风险因素，用不同方法估算的资源量，其可靠程度差别甚大，所以风险值也不同。

2．单位储量的综合勘探成本

单位储量的综合勘探成本是指由某种类型的石油资源，经过勘探获得单位地质储量（如100Mt、10Kt、1t等）的总投资费用。那么，第i种油气资源的单位储量综合勘探成本C_i的计算公式如下：

$$C_i=S_iH_iN_i/(K_{1i}K_{2i}) \quad (2\text{-}4\text{-}1)$$

式中 i——表示第i种储量；

C——单位储量的综合勘探成本；

S——每m钻井进尺的勘探成本；

H——探区中的探井平均井深；

N——获得单位储量所需要的探井数；

K_1——储量探井数与全部探井数的比值；

K_2——钻探费用与全部勘探费用的比值。

其中K_1的储量探井数是指含油面积内的探井数，而全部探井数中包括区域探井及落空探井。K_2的全部勘探费用中除钻探费用外，还包括物探费用（主要是地震勘探费用）及其他费用。

如果探区的费用是按勘探项目分别结算时，单位储量的综合勘探成本C_i可按下面的公式计算。

$$C_i=P_{1i}+P_{2i}+P_{3i} \quad (3\text{-}4\text{-}2)$$

式中 P_1——单位储量的钻探费用；

P_2——单位储量的物探费用。

P_3——单位储量的其他方面费用。

3．勘探方案实施期间探区逐年的计划产量以及已投入开发油田的预计产能

如果探区中已有n个油田投入开发，而勘探方案是从第k年开始实施，其实施期间的总年数为m年，探区在m年内预计的累计产量差额为Δq，则

$$\Delta q=\sum_{i=k}^{k+m-1}\Delta q_i=\sum_{i=k}^{k+m-1}(\hat{q}_i-\sum_{j=1}^{n}q_{ij}) \quad (3\text{-}4\text{-}3)$$

式中 Δq——探区在m年内预计的累计产量差额；

Δq_i——探区第i年预计的产量差额；

$\hat{q}_i$——探区第i年的计划产量指标；

q_{ij}——已投入开发的第j个油田第i年的预测产量。

探区中已投入开发的油田产量是n个油田的产量之和。其中未达到产量高峰的油田产量在以后的一段时间内可能上升，而过了产量高峰的油田产量将要逐渐下降。为了使探区的产量上升并且达到计划的产量指标，必须用新找到的油田弥补累计产量的差额Δq。

勘探方案实施的最后一年，即第$k+m-1$年的原油年产量记为q_{k+m-1}，当时的储采比按a%考虑时，设第$k+m-1$年剩余的可采储量为Q_{k+m-1}，则

$$Q_{k+m-1}=q_{k+m-1}/a\% \quad (3\text{-}4\text{-}4)$$

如果全探区地质储量的采收率按b%计算，则在勘探方案实施期间起码应当找到的地质储量Q为

$$Q=(\Delta q+Q_{k+m-1})/b\% \quad (3\text{-}4\text{-}5)$$

4. 探区中各分区之间的产量分配

探区中各分区之间的产量分配是指勘探方案实施期间，分区间的逐年计划产量与已投入开发油田的对应年份预测产量之差的累计和之间的比例。如果探区中共有s个分区，那么其a分区与b分区之间的产量比例k_{ab}为

$$k_{ab}=\sum_{i=k}^{k+m-1}(\hat{q}_{ai}-q_{ai})/\sum_{i=k}^{k+m-1}(\hat{q}_{bi}-q_{bi}) \qquad (3\text{-}4\text{-}6)$$

$$(a、b=1, 2, \cdots, s; \ a\neq b)$$

式中 k_{ab}——a分区与b分区之间的产量比例；

$\hat{q}_{ai}$、$\hat{q}_{bi}$——分别为a、b分区第i年的计划产量；

q_{ai}、q_{bi}——分别为a、b分区中已投入开发油田的第i年预测产量。

5. 各种储量类型之间的产量分配

按原油的质量可将储量划分为很多种类型，诸如构造型储量、低渗低产型储量、稠油型储量等。如果探区中有p种储量类型，那么，类型c与类型d之间的产量比例为k_{cd}，则

$$k_{cd}=\sum_{i=k}^{k+m-1}(\hat{q}_{ci}-q_{ci})/\sum_{i=k}^{k+m-1}(\hat{q}_{di}-q_{di}) \qquad (3\text{-}4\text{-}7)$$

$$(c、d=1, 2, \cdots, p; \ c\neq d)$$

式中 k_{cd}——类型c与类型d储量之间的产量比例；

$\hat{q}_{ci}$、$\hat{q}_{di}$——分别为类型c、d储量第i年的计划产量；

q_{ci}、q_{di}——分别为类型c、d储量的已投入开发油田第i年预测产量。

制定一个探区的勘探方案除涉及上述几项因素外，还可能需要考虑更多的因素，例如现有勘探技术所能承担的地质条件及地面条件的难度；探区中各分区的人文、经济情况；勘探方案实施期间可以动用的投资费用及勘探设备；石油价格的今后波动及勘探物资的今后比价等等。

二、线性规划模型

线性规划是运筹学的一个重要分支，自从1947年丹捷格（G.B.Dantzig）提出求解线性规划问题的单纯形法之后，线性规划在理论上已趋于成熟，实际应用日益广泛。特别是现在可以借助电子计算机来处理数量众多的约束条件与变量的大规模线性规划以后，它的适用领域更加广泛了。

线性规划问题都具有如下三个特征：

（1）可以用一组n个未知数（x_1，$x_2\cdots$，x_n）表示所研究问题的一个方案，这n个未知数的一组定值就代表一个具体方案。在实际制定方案时，这些未知数应当是非负的。

（2）存在一定的限制条件，这些条件称为约束条件，而约束条件可用一组m个线性等式或不等式表示。

（3）要有一个目标要求，并且这个目标可以表示为一组未知数的线性函数，这一线性函数称为目标函数。按所研究问题的特点，可要求目标函数达到最大或最小。

因此，线性规划问题可用如下数学模型表示：

（1）目标函数为

$$\text{Max}(\text{Min})\ z=c_1x_1+c_2x_2+\cdots+c_nx_n \qquad (3\text{-}4\text{-}8)$$

（2）约束条件为

$$\begin{cases} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n\leqslant(=,\ \geqslant)b_1 \\ a_{21}x_1+a_{22}x_2+\cdots+a_{2n}x_n\leqslant(=,\ \geqslant)b_2 \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots \\ a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n\leqslant(=,\ \geqslant)b_m \end{cases} \tag{3-4-9}$$

（3）非负条件

$$x_1,\ x_2,\ \cdots,\ x_n\geqslant 0 \tag{3-4-10}$$

为了书写方便，线性规划的数学模型可以缩写为如下形式

$$\text{Max(Min)}\{z=CX\leqslant(=,\ \geqslant)B,\ X\geqslant 0\} \tag{3-4-11}$$

其中：

$$X=\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}_{n\times 1}$$

$$C=\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}^T_{n\times 1}=[c_1c_2\cdots c_n]_{1\times n}$$

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}_{m\times n}$$

$$B=\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}_{m\times 1}$$

为了便于计算，上述的线性规划模型应当化为如下标准形式：

$$\text{Min}z=c_1x_1+c_2x_2+\cdots+c_nx_n \tag{3-4-12}$$

$$\begin{cases} a_{11}x_1+a_{12}x_2+\cdots+a_{1n}x_n=b_1 \\ a_{21}x_1+a_{22}x_2+\cdots+a_{1n}x_n=b_2 \\ \cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots\quad\cdots \\ a_{m1}x_1+a_{m2}x_2+\cdots+a_{mn}x_n=b_m \end{cases} \tag{3-4-13}$$

$$x_1,\ x_2,\ \cdots,\ x_r\geqslant 0 \tag{3-4-14}$$

这种标准形式亦可缩写为

$$\text{Min}\{z=CX\mid AX=B,\ X\geqslant 0\} \tag{3-4-15}$$

如果所建立的线性规划模型与标准型（3-4-15）式不同，可按下面方法进行变换：

（1）若要求$z=CX$的极大值，可令$z'=-CX$，然后求z'的极小值，如果z'有极小值存在，则z亦有极大值，且等于z'的极小值。

（2）若存在某个$b_i<0$（$i=1, 2, \cdots, n$），可将第i个方程两边乘以-1，就把b_i化成大于等于0。

（3）若约束条件中存在不等式关系，可分别按下面两种情况处理：

①当第i个约束条件为

$$a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_n\geqslant b_i$$

时，可引进剩余变量$x_{n+i}\geqslant 0$，使不等式化为下列等式

$$a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_i-x_{n+i}=b_i$$

②当第i个约束条件为

$$a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_n\leqslant b_i$$

时，可引进松弛变量$x_{n+i}\geqslant 0$，使不等式化为下列等式

$$a_{i1}x_1+a_{i2}x_2+\cdots+a_{in}x_n+x_{n+i}=b_i$$

通过上述变换就可以把一个线性规划问题转化为（3-4-15）式表示的标准型。剩下的问题就是该标准型的数学解了，目前通常采用改进的单纯形法。

三、算　　例

为了说明如何用线性规划方法制定勘探方案，这里以一个假想探区制定1986～1990年的五年勘探方案为例。

该探区为一个完整的沉积盆地，大部分面积在陆地上，小部分为浅海水域。陆地上根据地震资料解释可分为三个地质坳陷，即北部坳陷、南部坳陷、西部坳陷。由于海域地区投入的勘探工作量较少，地质结构不太清楚可暂切作为一个地质分区。该盆地经以往勘探已探明一批油田，并已投入开发。

根据以往的研究，陆地上三个地质坳陷的石油资源按油藏类型大体分为三类，即构造型资源、潜山型资源、岩性型资源。由于海域部分勘探成本较高，制定勘探方案时仅考虑容易勘探的构造型资源。

经过风险分析后的石油资源量预测结果见表3-4-1。不同类型的资源量，经过勘探后升级为可采储量的勘探成本是不一样的。根据上述三个坳陷及一个海域中预测的各类型资源量所处的地面条件、地质条件测算，获得每100Mt地质储量的勘探费用也列入表中。表中的x_1，x_2，…，x_{10}为石油资源类型。

制定该勘探方案的要求是1986～1990年期间探区的石油产量逐年略有增长。如果探区每年的计划产量为$\hat{q}$，老油田的预测产量为q，则第i年的新增产量应为$\Delta\hat{q}_i=\hat{q}_i-q_i$。那么，五年间累计的新增产量$\Delta q$应为

$$\Delta q=\sum_{i=1986}^{1990}(\hat{q}_i-q_i)=\sum_{i=1986}^{1990}\Delta q_i$$

为了保证探区在1990年以后能长期稳产，则1990年底尚需有足够的可采储量，若1990年底的储采比按a%考虑时，1990年底的剩余可采储量应为

$$Q_{1990}=q_{1990}/a\%$$

根据以上计算，在勘探方案实施期间起码应当找到的可采储量为

$Q=\Delta q+Q_{1990}$

经过计算Q等于550Mt。

表3-4-1　石油资源量及勘探成本数据表

分区名称	资源类型	风险分析后的资源量预测值（100Mt）	单位地质储量的勘探成本（亿元/100Mt）
北部坳陷	构造型x_1	1	2.5
	潜山型x_2	2	4
	岩性型x_3	2.7	8
西部坳陷	构造型x_4	1.5	3
	潜山型x_5	0.6	5
	岩性型x_6	2.2	9
南部坳陷	构造型x_7	0.2	4
	潜山型x_8	0.1	6
	岩性型x_9	0.1	10
海域地区	构造型x_{10}	4	30

关于探区中北部坳陷、西部坳陷、南部坳陷、海域地区之间的新增可采储量的比例，经过有关人员商定后认为1:0.5:0.05:0.2较为合适，即

$$(x_1+x_2+x_3):(x_4+x_5+x_6):(x_7+x_8+x_9):x_{10}$$
$$=1:0.5:0.05:0.2$$

探区中构造型、潜山型、岩性型资源之间新增可采储量之间的比例为1:0.5:0.5较为合适，即

$$(x_1+x_4+x_7+x_{10}):(x_2+x_5+x_8):(x_3+x_6+x_9)$$
$$=1:0.5:0.5$$

鉴于以上分析讨论，可以得到如下目标函数及约束条件：

目标函数为

$$\text{Min}z=2.5x_1+4x_2+8x_3+3x_4+5x_5+9x_6+4x_7+6x_8+10x_9+30x_{10}$$

式中　z——探区1986～1990年期间的总勘探成本（亿元），要求z为最小，亦即投资最少；

x_1、x_2、x_3——分别为北部坳陷新增的构造型、潜山型、岩性型可采储量（$\times10^8$t）；

x_4、x_5、x_6——分别为西部坳陷新增的构造型、潜山型、岩性型可采储量（$\times10^8$t）；

x_7、x_8、x_9——分别为南部坳陷新增的构造型、潜山型、岩性型可采储量（$\times10^8$t）；

x_{10}——海域地区新增的构造型可采储量（$\times10^8$t）。

约束条件如下

$$
\begin{cases}
\sum_{i=1}^{10} x_i \geqslant 5.5 \\
0.5\sum_{i=1}^{3} x_i - \sum_{i=4}^{5} x_i = 0 \\
0.05\sum_{i=1}^{3} x_i - \sum_{i=6}^{9} x_i = 0 \\
0.2\sum_{i=1}^{3} x_i - x_{10} = 0 \\
0.5(x_1 + x_4 + x_7 + x_{10}) - (x_2 + x_5 + x_8) = 0 \\
0.5(x_1 + x_4 + x_7 + x_{10}) - (x_3 + x_6 + x_9) = 0 \\
x_1 \leqslant 1 \\
x_2 \leqslant 2 \\
x_3 \leqslant 2.7 \\
x_4 \leqslant 1.5 \\
x_5 \leqslant 0.5 \\
x_6 \leqslant 2.2 \\
x_7 \leqslant 0.2 \\
x_8 \leqslant 0.1 \\
x_9 \leqslant 0.1 \\
x_{10} \leqslant 4
\end{cases}
$$

非负条件如下

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10} \geqslant 0$$

上述约束条件中，第一个约束条件的含义是探区总计新增的可采储量应大于或等于550

表3-4-2 新增可采储量数据表

分区名称	资 源 类 型	新增储量（100Mt）	合计储量（100Mt）
北部坳陷	构造型x_1 潜山型x_2 岩性型x_3	0.46 1.30 1.3	3.14
西部坳陷	构造型x_4 潜山型x_5 岩性型x_6	1.50 0.07 0	1.57
南部坳陷	构造型x_7 潜山型x_8 岩性型x_9	0.16 0 0	0.16
海域地区	构造型x_{10}	0.63	0.63
			5.50

Mt。

第2至第4个约束条件的含义是观北部、西部、南部、海域各坳陷的新增可采储量之间的比例应为1:0.5:0.05:0.2。

第5至第6条约束条件的含义是构造型、潜山型、岩性型新增储量之间的比例应为1:0.5:0.5。

第7至第16个约束条件的含义是各种类型新增的可采储量不能超过已预测出的资源量。而非负条件要求新增可采储量的数值为非负的。

上述模型化为标准型以后，经过计算得到如下结果。

$$z=41.72\text{亿元}$$

即探区在1986～1990年期间的计划投资应为41.72亿元，并且这是满足上述约束条件下的最少投资数额。探区的新增储量见表3-4-2。

第二节　石油勘探的决策分析

勘探决策是石油地质勘探过程中普遍存在的一种活动。它是为了解决当前或未来勘探过程中可能发生的情况而选择最佳勘探方案的一种过程。决策分析贯穿勘探管理的全过程，管理本身就是决策。例如，面对一个新探区，如何布置地震测线，首先钻探哪个圈闭，如何选择井位，钻探出油后如何调整勘探方案等等一系列问题，都需要有关的各级管理人员，特别是有裁决权的领导人及时作出决策。

勘探决策的效果有好有坏，决策是否正确，是否合理，小则关系到能否达到勘探的预期目的，大则可能涉及巨大的经费投资，因决策失误将会贻误时机而影响国家的经济建设。可见，勘探决策是否正确关系重大。为此，决策者必须具有科学的态度，并且掌握有关的决策原理和方法。

石油地质勘探过程中，经常会面临几种不同的客观条件，又会有几种可供选择的勘探方案。在这种情况下，勘探工作的管理人员必须面对不同的客观条件，在几种可以选择的方案中选出最优勘探方案加以实施，这就是勘探决策问题。

衡量勘探决策是否合理、是否正确的标准是勘探后的经济效益，而经济效益通常是以货币益损值表示。

勘探工作面临的几种客观条件，称为自然状态，简称状态。自然条件是客观条件，是不以人的主观意志为转移的，所以可称为不可控制因素。例如，一个探区中有3个地质凹陷，每个凹陷的含油气地质条件完全是由地质历史的演变所自然形成的。但是，勘探方案却是人为制定的，即如何进行勘探，选定何种部署方案是可控制因素，完全取决于决策者的主观选择。

勘探决策中的基本表达式为

$$a=F(A_i,\ \theta_j)$$

$$(i=1,\ 2,\ \cdots,\ m)\ (j=1,\ 2,\ \cdots n)$$

式中　a——益损值；

A_i——决策者的可控制因素，即勘探方案，把A_i当作变量看待时，可称作决策变量；

θ_j——决策者的不可控制因素，即自然状态，把θ_j当作变量看待时，称作状态变量。

从另一个角度看，θ_j为未确定因素，A_i表示决策者为应付出现状态θ_j所采取的决策方案；它们之间的关系可用如下益损矩阵表示。

$$
\begin{array}{cc}
 & \text{自然状态} \\
 & \begin{array}{cccc} \theta_1 & \theta_2 & \cdots & \theta_n \end{array} \\
\begin{array}{c} \text{勘} \ A_1 \\ \text{探} \ A_2 \\ \text{方} \ \vdots \\ \text{案} \ A_m \end{array} &
\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}
\end{array}
$$

通常勘探决策可按其问题的性质不同，分为确定型、风险型、不确定型三类。

一、确定型勘探决策

所谓确定型勘探决策应具备如下四个条件：

（1）存在决策人希望达到的一个明确的勘探前景，即预期的勘探效益。

（2）只存在一种确定的状态。

（3）可以制定决策人选择的两种或两种以上的勘探方案。

（4）不同勘探方案的效果可以测算，即方案的益损值可以计算出来。

例如，一个探区经过初步勘探，根据钻井及地震资料解释已控制20km²的含油面积，这是唯一的一个自然状态。根据石油地质勘探人员的认真分析，如果用两台钻机继续钻探，预计一年内可以拿下50Mt的石油地质储量。这是决策人希望达到的一个收益较大的勘探目标。如按30％采收率计算，可以采出15Mt原油。(若原油价格按120元/t计算，经济收益为18亿元。)如果不继续勘探，则拿不到地质储量，仅能根据20km²含油面积估算一下级别较低的石油地质储量，而没有直接经济效益。

通过这两个方案的比较，自然会选择第一个方案继续进行勘探。

确定型的勘探决策分析，看起来似乎非常简单，但在实际的勘探工作中并非如此。如果可供选择的勘探方案很多，为了搞清20km²含油面积的地质储量，往往会涉及多种勘探手段及相应的勘探方案。例如，上几个地震队，地震测线如何布置，上几台钻机，探井井位如何选定等都需要认真研究。因此，对于多种勘探方案常用最优化方法处理，例如用线性规划方法进行方案比较。

需要指出，确定型的石油勘探决策非常少见，绝大多数石油勘探决策都属于风险型或者不确定型勘探决策。

二、风险型勘探决策

风险型勘探决策也称随机型勘探决策，一般它具有如下五个条件：

（1）存在决策人希望达到的收益较大的勘探前景。

（2）有两个或两个以上的勘探方案可供决策人选择，最后只需选择一个方案。

（3）存在两个或两个以上不以决策人的主观意志为转移的自然状态。

（4）不同勘探方案的各种自然状态下的益损值可以计算出来，即可以建立益损矩阵。

（5）在几种不同的自然状态中，经过勘探出现哪种自然状态，决策人事先不能肯定；但是，各种自然状态出现的可能性，即其概率值，决策人可以预先估计或者预测出来。

对于风险型勘探决策问题，常用的计算方法有如下三种：

（1）最大可能法　某个事件，预计出现的概率越大，发生的可能性也就越大。因此，最大可能法就是选择出现概率最大的自然状态进行决策，而对其他的自然状态可以不管，这就使风险型决策问题变成确定型决策问题。现举例如下，见表3-4-3。

表3-4-3　风险型决策分析数据表

自然状态 / 勘探方案	θ_1	θ_2	θ_3
	$P(\theta_1)=0.3$	$P(\theta_2)=0.5$	$P(\theta_3)=0.2$
A_1	0Mt −500万元	20Mt 100000万元	80Mt 400000万元
A_2	1Mt 4500万元	10Mt 45000万元	50Mt 225000万元
A_3	0Mt −1000万元	50Mt 200000万元	100Mt 400000万元
A_4	2Mt 8000万元	40Mt 150000万元	80Mt 300000万元

表3-4-3中的A_1、A_2、A_3、A_4表示有4个探区，亦即存在4个可供决策者选择的勘探地区。而勘探任何一个探区均有三种可能的勘探结果，即可能出现θ_1、θ_2、θ_3三种自然状态中的一种。而三种自然状态的出现概率$p(\theta_1)$、$p(\theta_2)$、$p(\theta_3)$分别等于0.3、0.5、0.2。表3-4-3中的益损值给出了经过勘探后可能获得的可采储量及扣除投资后的净收入（或亏损），例如A_3方案出现θ_2时，可以获得可采储量50Mt，扣除投资后的净收入为20亿元。

对于这个算例，按最大可能法决策时，则要按自然状态θ_2考虑究竟选择哪个探区，这是因为$p(\theta_2)=0.5$，与$p(\theta_1)$、$p(\theta_3)$相比其出现的可能性最大。比较表3-4-3中在θ_2自然状态下的A_1、A_2、A_3、A_4的益损值，A_3探区的收益值最大，可获得50Mt可采储量，扣除勘探及油田建设投资后，可获纯利润20亿元。

需要指出，用最大可能法进行勘探决策时要注意，在几种可能的自然状态中，有一种自然状态出现的概率特别大时，决策的效果才好；如果n种自然状态中，各种状态出现的概率都不大，而且相互接近，则采用最大可能法进行决策时效果好坏的把握性不大，甚至会出现重大失误。

（2）期望值法　如果把每个勘探方案看成是离散型随机变量，那么，可以用离散型随机变量的数学期望来表示每个勘探方案的益损值。

设：$A=\{A_1, A_2, \cdots, A_m\}$

为所有可能的勘探方案集合，A_i（$i=1, 2, \cdots m$）是集合中的元素。

也可以把A看作是一个向量，即

$$A=(A_1, A_2, \cdots A_m)$$

可称A为勘探方案向量，A_i是它的分向量。

同理可设：$\theta=\{\theta_1, \theta_2, \cdots, \theta_n\}$

为所有的自然状态集合，θ_j（$j=1, 2, \cdots n$）是集合中的元素。

也可以把θ看作是一个向量，即

$$\theta=(\theta_1, \theta_2, \cdots \theta_n)$$

可称θ为自然状态向量，θ_j是它的分向量。

若自然状态θ_j的发生概率记作$p(\theta_j)=p_j$，则p可称为自然状态概率向量，即

$$p=(p(\theta_1), p(\theta_2), \cdots, p(\theta_n))=(p_1, p_2, \cdots, p_n)$$

显然有

$$\sum_{j=1}^{n} p_j=1$$

当采用勘探方案A_i时，出现自然状态θ_j的益损值可记为$a(A_i, \theta_j)$，简记为a_{ij}。

A_i的益损期望值为

$$E(A_i)=\sum_{j=1}^{n} p_j a_{ij} \tag{3-4-16}$$

令

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

A称为益损矩阵或风险矩阵。如果把$E(A)$称为A的期望值，则有

$$E(A)=\begin{pmatrix} E(A_1) \\ E(A_2) \\ \cdots \\ E(A_m) \end{pmatrix}=Ap^T=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix}=\begin{pmatrix} \sum_{j=1}^{n} p_j a_{1j} \\ \sum_{j=1}^{n} p_j a_{2j} \\ \cdots \quad \cdots \\ \sum_{j=1}^{n} p_j a_{mj} \end{pmatrix} \tag{3-4-17}$$

上式就是按期望值法进行决策时的计算公式。最后，可按（3-4-18）式选择勘探方案：

$$A_{max}=\max[E(A)] \tag{3-4-18}$$

根据表3-4-3给出的数据，按期望值法计算如下

$$E(A)=\begin{pmatrix} -500 & 100000 & 400000 \\ 4500 & 45000 & 225000 \\ -1000 & 200000 & 400000 \\ 8000 & 150000 & 300000 \end{pmatrix}\begin{pmatrix} 0.3 \\ 0.5 \\ 0.2 \end{pmatrix}=\begin{pmatrix} 129850 \\ 68850 \\ 179700 \\ 107400 \end{pmatrix}$$

$$A_{max}=\max[E(A)]=179700=A_3$$

即，应该选定第三个探区进行勘探。

（3）决策树法　决策树法因其过程似树形而得名，决策的过程犹如从树枝归结到树干。

按决策树法进行勘探决策时的步骤如下：

①首先画出决策树的分支结构。决策树实际上就是把勘探决策问题的各种可能发生情况及其所作的预测，用树形表示出来。画决策树的过程就是拟定并选择各种勘探方案的过程。

②预测勘探过程中可能发生的各种勘探结果的出现概率，即确定出现各种自然状态的概率。概率值的确定，可根据勘探人员的经验估计，或者根据含油气地质条件相似探区的资料推测。并且把概率值标在决策树图上的相应位置。

③最后计算益损期望值，从决策树的末梢向主干方向逐步计算。

表3-4-3中给出的数据，按决策树法进行勘探决策见图3-4-1。

图3-4-1中的□表示决策节点，归结到这里的分枝叫勘探方案分枝，分枝数就是可以选择的勘探方案A_i的数量。○称为勘探方案节点，它上面的数字表示勘探方案的效益值$E(A_i)$，归结到这里的分枝叫概率分枝，其上要标明自然状态的出现概率$p(\theta_i)$，分枝数就是可能出现的自然状态θ_i的数量。△叫作末梢点，标出的数值为相应自然状态下的益损值。

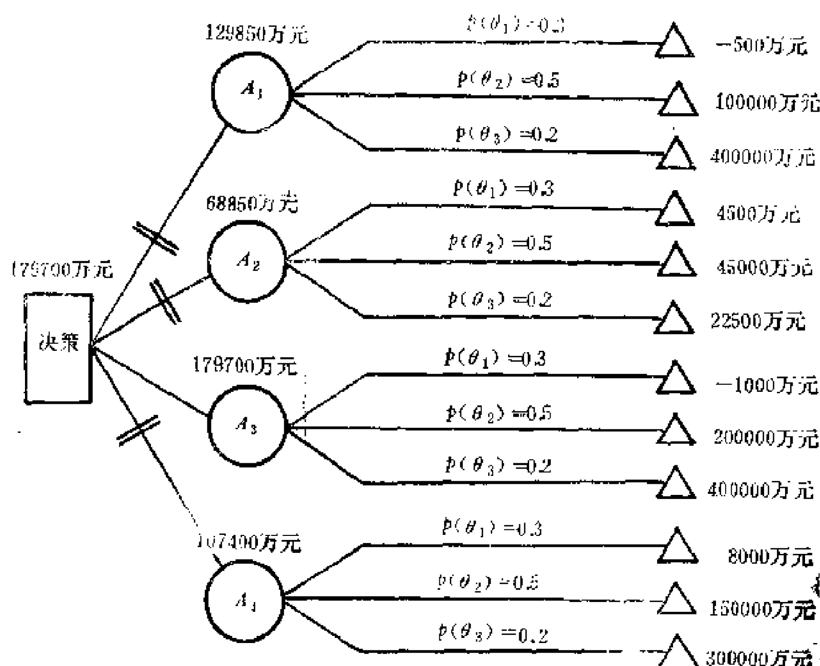


图3-4-1　单级决策树图

根据图3-4-1中各方案节点上的效益值加以比较，选择最大效益值179700万元作为决策方案，即选定A_3探区进行勘探，而其他三个勘探方案可以打上删除号||，称为剪枝方案。

上面介绍的决策树图中只有一个决策节点，因而称为单级决策。如果决策树图中有多个决策节点则叫多级决策。例如，某个探区中有一个重力高，此时有两种勘探方案可选择，其一是上地震队落实地下构造情况，其二是直接上钻井队进行钻探。如果上地震队，经过地震勘探后可能出现三种情况，即地下存在形态完整的局部构造、鼻状构造或没有构造。按重力高直接钻探，或者上地震队落实构造后再钻探，也都有三种可能，即打成高产井、低产井或

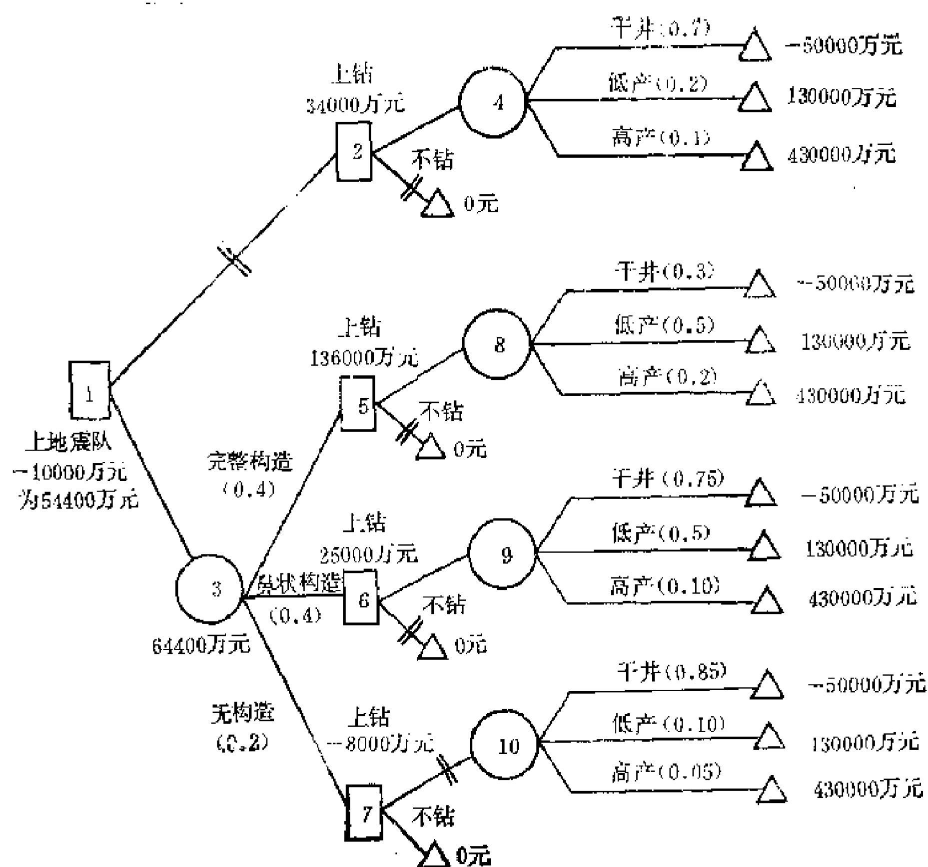


图3-4-2 多级决策树图

干井，见图3-4-2。

从图3-4-2看出，直接在重力高上打探井效益期望值为34000万元；上地震队进行地震勘探要支出10000万元，但是落实构造后再打探井的效益期望值为64400万元，减去地震队勘探费用支出后的效益期望值为54400万元。所以，最后选择上地震队落实构造后再打探井的勘探方案。

方案节点3为上地震队落实构造后再打深井，其效益期望值是由地震勘探后发现完整构造、鼻状构造、无构造的发生概率以及上钻打井或不打井的效益期望值所决定的，即

$$136000\times0.4+25000\times0.4+0\times0.2=64400\text{万元}$$

而构造落实后上钻或不上钻，决定于出现高产井、低产井、干井的概率及益损值。如果有完整构造，上钻时为方案节点8，效益期望值为

$$-50000\times0.3+130000\times0.5+430000\times0.2=136000\text{万元}$$

不上钻时益损值为0元。所以应当上钻，即采用决策节点5。

如果有鼻状构造，上钻时为方案节点9，期望值为

$$-50000\times0.75+130000\times0.15+430000\times0.1=25000\text{万元}$$

不上钻时益损值为0元。所以应当上钻，即采用决策节点6。

如果不存在构造，上钻时为方案节点10，期望值为

$$-50000\times0.85+130000\times0.1+430000\times0.05=-8000\text{万元}$$

不上钻时益损值为0元。所以不应当上钻，即采用决策节点7。

不上地震队而直接钻探重力高时为方案节点4，期望值为

$$-50000\times0.7+130000\times0.2+430000\times0.1=34000\text{万元}$$

不钻探重力高时益损值为0元。所以应当钻探，即采用决策节点2。

通过上面的算例说明，决策树法比较直观，使决策人能够按勘探程序逐步决策。特别是对复杂的多级决策问题，用决策树法便于石油地质人员集体讨论，使决策工作能够在集思广益的基础上，以科学的推理去周密思考勘探过程中的各种有关因素以及可能出现的各种情况。

三、不确定型勘探决策

不确定型勘探决策是风险勘探决策缺少第五个条件下的决策问题，也就是说，勘探过程中可能出现的各种自然状态的概率，事先无法估计。例如，在一个新的盆地进行早期勘探时，因为极度缺乏地质资料，并且没有可以借鉴的类比含油区，这种情况下进行决策即为不确定型的勘探决策问题。

例如在某个新探区，共有四种可供选择的勘探方案，即A_1、A_2、A_3、A_4。而每种勘探方案都有三种可能出现的勘探结果，即有自然状态θ_1、θ_2、θ_3，但是，出现θ_1、θ_2、θ_3的概率无法估计，见表3-4-4。

表3-4-4　不确定型勘探决策分析数据表（单位：万元）

自然状态 / 勘探方案	θ_1	θ_2	θ_3
A_1	−200	1000	1200
A_2	600	1900	1000
A_3	−500	400	2500
A_4	400	200	1600

表3-4-4中的数据是勘探益损值，就是扣除了勘探投资费用后的纯利润收入（或亏损费用）。

对于不确定型的勘探决策，有以下五种常用方法：

（1）乐观决策法　如果在某个探区勘探时，有m种可供选择的勘探方案，预计每种勘探方案可能出现n种自然状态，第i种勘探方案第j种自然状态的益损值为$a(A_i,\theta_j)=a_{ij}$。

乐观决策法的计算公式为

$$\max_{A}\{\max_{\theta}[a(A,\theta)]\} \tag{3-4-19}$$

即首先在益损矩阵中选出每行的最大效益值构成列向量，再从这个列向量中选出最大的效益值，该值对应的勘探方案就是选定的勘探方案。

按表3-4-4给出的数据，用乐观决策法进行勘探决策时有

$$\max_{\theta}[a(A,\theta)]$$

$$\begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{array}\begin{array}{c}\begin{array}{ccc}\theta_1 & \theta_2 & \theta_3\end{array}\\ \begin{pmatrix} -200 & 1000 & 1200 \\ 600 & 1900 & 1000 \\ -800 & 400 & 2500 \\ 400 & 200 & 1600 \end{pmatrix}\end{array} \qquad \begin{pmatrix} 1200 \\ 1000 \\ 2500 \\ 1600 \end{pmatrix}$$

$$\max_{A}\{\max_{\theta}[a(A,\theta)]\}=2500$$

即，按乐观决策法计算，应选择A_3方案进行勘探。

（2）悲观决策法　在探区的地质特征不太清楚的情况下，对勘探工作要持慎重态度，即认为勘探工作不会太顺利。因此悲观决策法是在所有可能的勘探方案中选择最不好的自然状态，再从最不好的自然状态中找出相对较好的勘探方案。

悲观决策法的计算公式为

$$\max_{A}\{\min_{\theta}[a(A,\theta)]\} \tag{3-4-20}$$

可见，悲观决策法是对勘探前景持慎重态度的、偏于保守的方法。

按表3-4-4给出的数据，用悲观决策法进行勘探决策时有

$$\min_{\theta}[a(A,\theta)]$$

$$\begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{array}\begin{array}{c}\begin{array}{ccc}\theta_1 & \theta_2 & \theta_3\end{array}\\ \begin{pmatrix} -200 & 1000 & 1200 \\ 600 & 1900 & 1000 \\ -800 & 400 & 2500 \\ 400 & 200 & 1600 \end{pmatrix}\end{array} \qquad \begin{pmatrix} -200 \\ 600 \\ -800 \\ 400 \end{pmatrix}$$

$$\max_{A}\{\min_{\theta}[a(A,\theta)]\}=600$$

即，按悲观决策法计算，应选择A_2方案进行勘探。

（3）估计系数法　估计系数法是对探区的勘探前景既不抱乐观态度，也不抱悲观态度，而是用［0，1］区间上的一个数值代表勘探前景，这个数值α称为估计系数。

估计系数法的计算公式为

$$\max_{A}\{\alpha\max_{\theta}[a(A,\theta)]+(1-\alpha)\min_{\theta}[a(A,\theta)]\} \tag{3-4-21}$$

如果$\alpha=1$时，则变为乐观决策法，即

$$\max_{A}\{\max_{\theta}[a(A,\theta)]\}$$

如果$\alpha=0$时，则变为悲观决策法，即

$$\max_{A}\{\min_{\theta}[a(A,\theta)]\}$$

现假定$\alpha=0.5$，则$1-\alpha=0.5$。按表3-4-4给出的数据，用估计系数法进行勘探决策时，要先从每个勘探方案中选出最大效益值乘以α，再选出最小的效益值乘以$(1-\alpha)$，二者相加的和代表估计系数为α时的效益值。再从这些新的效益值中找出最大效益值，其所对应的勘探方案为决策方案。

$$\alpha\max_{\theta}[a(A,\theta)]+(1-\alpha)\min_{\theta}[a(A,\theta)]$$

$$\begin{matrix} & \theta_1 & \theta_2 & \theta_3 \\ A_1 \\ A_2 \\ A_3 \\ A_4 \end{matrix}\begin{pmatrix} -200 & 1000 & 1200 \\ 600 & 1900 & 1000 \\ -800 & 400 & 2500 \\ 400 & 200 & 1600 \end{pmatrix} \quad \begin{pmatrix} 500 \\ 1250 \\ 850 \\ 900 \end{pmatrix}$$

$$\max_{A}\{\alpha\max_{\theta}[a(A,\theta)]+(1-\alpha)\min_{\theta}[a(A,\theta)]\}=1250$$

即，当$\alpha=0.5$时，按估计系数法计算应选择勘探方案A_2进行勘探。

如果取$\alpha=0.8$，$(1-\alpha)=0.2$时，计算后得出应选择勘探方案A_3进行勘探。

如果取$\alpha=0.2$，$(1-\alpha)=0.8$时，计算后得出应选择勘探方案A_2进行勘探。

可见，α取值不同，勘探决策结果可能不同。而如何确定α的取值，要根据石油地质勘探人员对勘探前景的估计。一般情况下，含油气地质条件较好时，则应把α值取大一些，反之，α值应取小些。

（4）等可能法　既然是不确定型勘探决策，就是说对各种可选择的勘探方案，出现哪种自然状态的概率是不能预先估计的，那么，是否可以认为出现各种自然状态的概率是相等的呢？等可能法就是以"一视同仁"的原则认定各种自然状态的出现概率是相等的。即，如果有n种可能出现的自然状态，则认为每种自然状态出现的概率为$\frac{1}{n}$，然后按风险型勘探决策中的期望值法进行决策即可。

按表3-4-4表给出的数据，用等可能法进行勘探决策时，每种勘探方案的三种自然状态出现的概率均定为$\frac{1}{3}$因而，有

$$E(A)=Ap^t=\begin{pmatrix} -200 & 1000 & 1200 \\ 600 & 1900 & 1000 \\ -800 & 400 & 2500 \\ 400 & 200 & 1600 \end{pmatrix}\begin{pmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{pmatrix}=\begin{pmatrix} 666.67 \\ 1166.67 \\ 700.00 \\ 733.33 \end{pmatrix}$$

$$A_{max}=\max[E(A)]=1166.67=A_2$$

即，按等可能法计算，应选择勘探方案A_2进行勘探。

（5）后悔值法　按决策方案进行勘探后，如果勘探效果并不理想，此时，决策者往往会感到后悔。"后悔值"方法的出发点，就是把每种自然状态中的最高效益值作为该状态下的理想目标。用该值减去其他方案的效益值，所得之差称为未达到理想目标的后悔值。具体作法是在益损值矩阵的每列元素中选出数值最大的元素作为理想目标，再用理想目标减去本列中各行的元素值，所得之差构成的矩阵称为"后悔值矩阵"。

按表3-4-4给出的数据，后悔值矩阵如下

$$
\begin{array}{c}
 \\
 \\
A_1 \\ A_2 \\ A_3 \\ A_4
\end{array}
\begin{array}{c}
\begin{array}{ccc} \theta_1 & \theta_2 & \theta_3 \end{array} \\
\begin{pmatrix} 800 & 900 & 1300 \\ 0 & 0 & 1500 \\ 1400 & 1500 & 0 \\ 200 & 1700 & 900 \end{pmatrix}
\end{array}
\qquad
\begin{array}{c}
\max\limits_{\theta}[a(A,\theta)] \\
\begin{pmatrix} 1300 \\ 1500 \\ 1500 \\ 1700 \end{pmatrix}
\end{array}
$$

$$\min_{A}\{\max_{\theta}[a(A,\theta)]\}=1300$$

后悔值矩阵建立后，先从矩阵的每行中选出最大值构成列向量，再从这一向量中选出最小的后悔值，这一后悔值所对应的勘探方案就是要选定的决策方案。计算公式为

$$\min_{A}\{\max_{\theta}[a(A,\theta)]\} \tag{3-4-22}$$

按表3-4-4给出的数据，用后悔值法进行勘探决策时，应选择勘探方案A_1进行勘探。

鉴于上述，不确定型勘探决策问题，往往会因采用不同计算方法，得到不同的决策结果，而且也难以判断哪种决策的效果好，哪种决策的效果不好。出现这种情况是由于各种勘探方案可能出现自然状态的概率难以估计所造成的。

可见，进行不确定型勘探决策时，采用何种计算方法，完全取决于勘探工作中决策人所持的态度。对于勘探前景持乐观态度者可采用乐观决策法；持悲观态度者可采用悲观决策法；持中间状态度者可用估计系数法；持一视同仁态度者可用等可能法；对决策失误持后悔态度者可采用后悔值法。

第三节　效 用 理 论

前面讲述经济评价及勘探决策方法时，均以货币效益期望值作为衡量、选择勘探方案的标准。这种作法是否合理，值得讨论。

需要指出，同样一笔货币量，在不同场合下，在人们的主观上可能具有不同的价值。经济学家通常以"效用"这个概念去衡量人们对同等货币量在主观上的价值，这就是货币的效用值概念，效用理论也称偏爱理论。

例如，决策者面临这样一种选择，而且只允许有一次选择机会：第一种方案是毫无条件地发给决策者100元；第二种方案是采用抽签办法，抽中或抽不中的概率各为50%，抽中时发给决策者300元，抽不中则不发钱。试问，决策者应选择哪个方案？对于大多数决策者，特别是经济上不富裕的决策者，宁愿稳得100元，尽管采用抽签方案时的货币期望值为

$$300\times0.5+0\times0.5=150\text{元}$$

比采用第一方案多50元，也不愿意去冒什么也得不到的风险。

但是，如果第二种方案的发钱数量由300元增加到500元，而其他条件不变，试问决策者又如何选择？此时，决策者也可能认为值得冒一次风险，而采用抽签办法。

如果第二种方案的发钱数量还是300元，但是，抽中的概率由50%增加到75%，而抽不中的概率由50%下降到25%，此时，决策者也可能宁愿冒点风险而采用抽签办法。

通过这个例子说明，稳得100元，有50%的机会得到500元，有75%的机会得到300元这

三种情况对同一个决策者来说具有等价性。

又如，在某一海洋大陆架钻探石油，预计钻获油气的可能性极大，但是，钻探投资也相当大。这对于小石油公司来说，虽然勘探前景很好，但要冒风险，如果勘探工作不顺手，有可能使钻探费用超过公司的支付能力，从而导致公司破产。因此，对于一个小石油公司来说，它宁愿放弃这次诱人的海上勘探，也不愿冒经济破产的危险。但是，对于一个资金雄厚的大石油公司来说，它绝不会放弃这次海上勘探，因为即使勘探失败，对于大公司也没有多大的影响。

上面的两个例子说明，单凭货币期望值作为经济评价或勘探决策的标准并非完全合理，有时要用另一种标准，即效用期望值。

对于石油勘探的同一决策者来说，在不同勘探风险情况下，同等货币值具有不同的效用值，而在相同勘探风险情况下，由于不同的决策者对于勘探风险的态度不同，同一期望值也可能具有不同的效用值。

效用值通常以［0，1］闭区间上的数值表示，1表示最大效用值，0表示最小效用值。

一、效用曲线及其类型

以直角坐标系的横坐标表示益损值，纵坐标表示效用值，决策者对待风险态度的变化曲线，称作效用曲线。一般来说，不同决策人的效用曲线是不同的。

效用曲线可用心理试验法绘制。例如在某个沉积盆地进行石油勘探，如果有两种勘探方案可供选择，第一种勘探方案预计有50％的机会净收入100万元，有50％的机会损失50万元；第二种勘探方案预计有100％的机会净收入10万元。在此情况下，试问决策者采用哪个方案？这里规定100万元的效用值为1，－50万元的效用值为0。如果这个决策者是个不愿冒风险的人，他认为采用第二方案比较稳妥，可以稳得10万元收入。

如果第二方案预计100％的机会只能净收入5万元，再问决策者采用哪个方案？此时，决策者认为采用第一方案或第二方案都可以。这就是说，决策人认为5万元的效用值与第一方案的效用值相等，即：$0.5\times1+0.5\times0=0.5$。所以，5万元的效用值为0.5，这样便得到效用曲线上的第一个点。

接下来，缩小益损值区间，即有50％的机会净收入100万元，有50％的机会净收入25万元时，重复上述询问过程，经过多次问答后确定50万元的效用值与这个方案的效用值相等，即：$0.5\times1+0.5\times0.5=0.75$。所以，50万元与效用值0.75相对应，这样就得到效用曲线上的第二个点。

同样再询问益损值的另一区间，即有50％的机会净收入25万元，有50％的机会损失50万元，经过多次询问后确定－25万元与效用值0.25相对应，即：$0.5\times0.5+0.5\times0=0.25$。所以，又得到效用曲线上的第三个点。

如此反复进行，则可以得到很多个益损值与效用值的对应点。把这些点连接起来，就可以得到这个决策者的效用曲线，见图3-4-3。

效用曲线的基本类型有三种，见图3-4-4。

1．保守型效用曲线

保守型效用曲线为图3-4-4中的曲线①，这种效用曲线所代表的决策人，其特点是认为肯定得到的某一效益值对应的效用值，大于带有风险的相等效益值对应的效用值。也就是

说，这种类型的决策人对于利益反应比较迟钝，而对损失比较敏感，是不求大利、回避风险、小心谨慎的保守型决策人。如果这种效用曲线的函数y_1有二阶导数，则有：

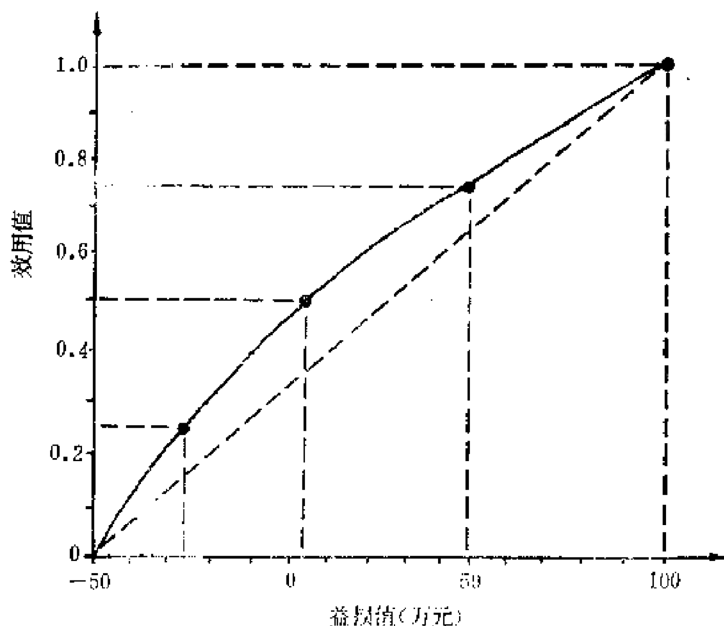


图3-4-3　效用曲线的绘制过程

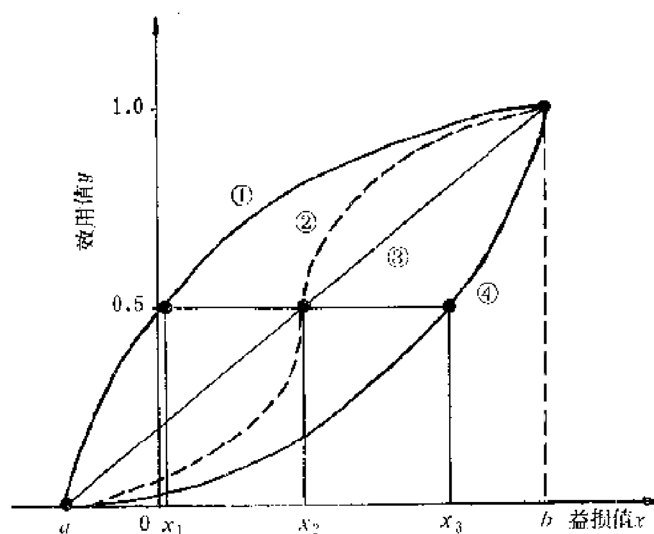


图3-4-4　效用曲线的类型

$$\frac{d^2y_1}{dx^2}<0$$

2．中间型效用曲线

中间型效用曲线为图3-4-4中的曲线②，这种效用曲线所代表的决策人，其特点是对风险不关心，认为益损值的效用值与益损值成正比。如果这种效用曲线的函数y_2有二阶导数，则有：

$$\frac{d^2y_2}{dx^2}=0$$

3. 冒险型效用曲线

冒险型效用曲线为图3-4-4中的曲线③，这种效用曲线所代表的决策人，其特点是认为肯定得到的某一效益值对应的效用值，小于带有风险的相等效益值对应的效用值。也就是说，这种类型的决策人对于损失反应比较迟钝，而对利益反应比较敏感，是谋求大利、敢冒风险的进取型决策人。如果这种效用曲线的函数y_3有二阶导数，则有：

$$\frac{d^2y_3}{dx^2}>0$$

除上述三种基本类型的效用曲线外，还可以由这三种类型组成多种不同的效用曲线。例如，在小益损值情况下决策人敢冒风险，而在大益损值情况下决策人又迴避风险，这种效用曲线为图3-4-4中的曲线④。

二、效用曲线的应用

在介绍效用理论之前，经济评价与勘探决策均以益损期望值作为衡量标准。如果考虑到决策人对勘探风险所持的态度，则应改用益损期望值的效用值作为衡量标准。

例如，在某个二级构造带进行整体解剖勘探，有两种可以选择的勘探方案，每种勘探方案预计有两种可能出现的自然状态，每种自然状态发生的概率均为50%。按决策树法计算时如图3-4-5所示。

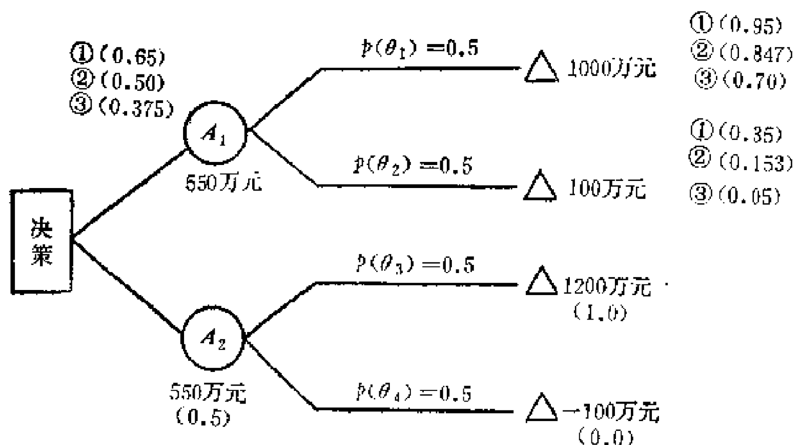


图3-4-5 效用值决策树图

对这一勘探问题进行决策时，在讨论中有三种意见，各种意见的效用曲线见图3-4-6中的曲线①、②、③。

图3-4-5中的勘探方案A_1和A_2的效益值均为550万元。所以，选用A_1或A_2勘探方案均可以。但是，按效用值决策时，则结果完全不相同。按图3-4-6中的三种效用曲线，1200万元为效益最大值，其效用值为1.0；－100万元为效益最小值，效用值为0。1000万元①、②、③三种效用曲线的效用值分别为0.95、0.846、0.7；100万元①、②、③三种效用曲线的效用值分别为0.350、0.153、0.050。

经计算方案A_1的550万元效益期望值的①、②、③三种效用曲线的效用值为0.65、0.5，0.375。即：

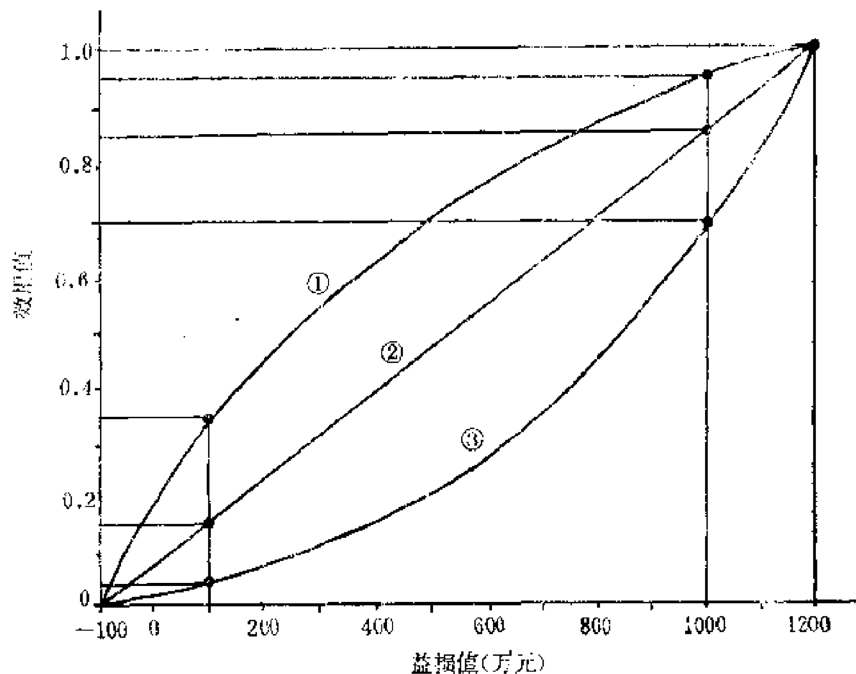


图3-4-6 效用值曲线

$$0.950\times0.5+0.350\times0.5=0.650$$

$$0.847\times0.5+0.153\times0.5=0.500$$

$$0.700\times0.5+0.050\times0.5=0.375$$

可见，按效用曲线①决策时，要选用A_1方案，因为A_1方案的效用值为0.65，大于A_2方案的效用值0.5。这是因为曲线①代表的决策者是个迴避风险的人，按A_1方案勘探至少也能稳得100万元的收益。

按效用曲线②决策时，选用A_1方案或A_2方案均可，因为A_1、A_2方案的效用值均为0.5。这是因为曲线②代表的决策者对风险不关心，认为效用值与益损值成正比。

按效用曲线③决策时，要选用A_2方案，因为A_2方案的效用值为0.5，大于A_1方案的效用值0.375。这是因为效用曲线③所代表的决策者敢冒风险，宁愿冒损失100万元的风险，也要争取得到1200万元。

又如，表3-4-4中给出的数据为一不确定型勘探决策问题，按等可能法计算时，A_1、A_2、A_3、A_4各方案的效益期望值分别为666.67万元、1166.67万元、700万元、733.33万元。四个方案按效益期望值大小排列时，顺序为A_2、A_4、A_3、A_1。

如果决策人的效用曲线为图3-4-7，按等可能法计算时，益损矩阵A要换成矩阵B，矩阵B中的元素由效用值构成，即

$$E(B)=Bp^t=\begin{pmatrix}0.05 & 0.70 & 0.79\\ 0.25 & 0.94 & 0.70\\ 0.00 & 0.17 & 1.00\\ 0.17 & 0.12 & 0.90\end{pmatrix}\begin{pmatrix}\frac{1}{3}\\ \frac{1}{3}\\ \frac{1}{3}\end{pmatrix}=\begin{pmatrix}0.1533\\ 0.6300\\ 0.3900\\ 0.3967\end{pmatrix}$$

经计算四个方案按效用值大小排列的顺序为A_2、A_4、A_1、A_3。

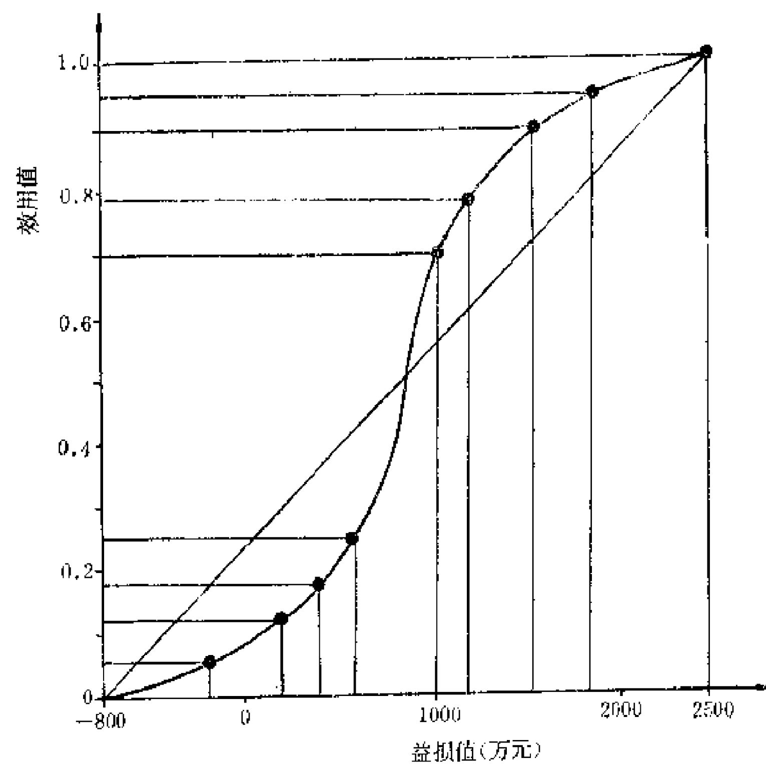


图3-4-7　决策人效用值曲线

第五章　石油资源评价的专家系统

近十年来，人工智能发展迅速，其中应用最广泛的是专家系统。目前，世界上许多国家都很重视专家系统的研制工作。据不完全统计，已开发的专家系统已达44种之多，已涉及到医学诊断、地质勘探、遗传工程、化学合成、公司管理、法律案件、军事战略等诸多方面。其中，美国、日本、法国在开发专家系统方面处于领先地位，并且已取得明显的社会效益。

专家系统技术近年在石油地质勘探中也得到迅速发展与应用。例如：

1．沉积盆地分类专家系统（MILLER，美国）

该系统提供了全世界600多个沉积盆地的资料，并以板块构造理论对这些盆地进行了分类，它可用于盆地石油资源的早期评价。

2．生油岩预测知识库系统（FOWLER，MICHAEL，美国）

该系统可以利用古地理、古气候等因素进行生油岩评价，使初学者好象专家一样进行复杂的逻辑推理，并且能够达到评价探区生油潜力之目的。

3．井间地层对比专家系统（SAITO，KATSUEI等，日本）

该系统可根据输入的倾角测井数据、地震勘探数据、岩相分析资料，使用知识库中的地层对比知识，首先确定标志层，然后进行砂岩对比，预测砂体几何形状和地层结构等，其对比结果存在数据库中。

4．古气候模拟专家系统（SCOTES，CHRISTOPHER，美国）

该系统可以对任何一个具体的古地理环境进行模拟古气候并绘制大气压的等值线图，再用压力梯度确定风向与古海岸线，最后提供潜在生油岩的分布线索。

5．用于地震地层学解释的专家系统（FANG，A.W.SHULTE，美国）

该系统需要输入地震剖面，剖面中的信息是通过语言变量来描述一个或多个特征，并且使用了模糊数学方法计算语言描述的不确定性，即可信度。解释结果包括大陆架的界限和倾斜度、盆地充满程度与地层的褶皱情况等。

6．测井曲线对比专家系统（KUO，STAYTMAN，美国）

该系统可以模仿地质学家的思路，首先判别单个岩相，继而详细刻划每个地层的特征，最后与另一口井相同的层位进行对比，对于对比结果可以通过原始层位框架检查，自动改善对比结果。

我国研究专家系统起步较晚，开始于本世纪80年代初。最早出现于医学诊断中，例如北京中医学院的“诊治内外科疾病系统”，上海计算所的“中医内科诊断系统”。80年代中期已扩展到其他领域。近几年，在石油勘探领域也先后研制出一些专家系统，例如：

1．石油测井解释专家系统（OWLI）

该系统是基于石油测井专家们的综合解释知识和经验的计算机咨询系统，也是一个决策系统，即模拟测井专家的推理思维，分析处理地球物理、钻井、测井等广义数据，给出地层的含油、气、水的评价，并能提供有关的地质参数。该系统是由华北油田勘探开发研究院与中国科学院自动化研究所共同研制的，包括模拟学习、单井解释、知识获取、推理、解

答、传播、修改、控制策略、知识库、数据库、词典库总共10个部分。知识库中包含4个标准画面，400多条推理判断规则以及50多个测井解释常用定量计算方程或模型。该系统是用GCLISP语言编制程序，可在IBM-PC/XT、AT微型机上操作使用。经过对华北油田110口井的实际解释表明，该系统具有较强的综合解释能力和较好的解释精度。

2. **预测地质圈闭含油气状况的专家系统**

该系统是把系统工程方法与专家系统方法有机结合起来，用来预测地质圈闭含油气状况的综合评价系统，由中国地质大学与吉林大学共同研制完成。这一系统主要由知识库、数据库、推理机、解释、学习五个部分组成。知识库中的知识表示采用分级分领域的框架式结构，其中一级知识是预测地质圈闭含油气状况的知识概念，包括概念特征、概念类型等；二级知识是在一级知识基础上构造出来的，用于描述概念之间的相互依赖关系；三级知识是对依赖关系进行具体的定量、定性、半定量的描述，三级知识可称为广义函数体，包括模糊关系、概率关系、灰度关系、线性关系、映射关系以及经验表示与逻辑推理等；四级知识是根据前三级知识的内容进行评价、选择与控制，四级知识可称控制关系，是建造专家系统推理机的基础；五级知识是对前四级知识在计算机中运行结果进行评价，对不确定性的传播进行描述。

3. **油气资源评价专家系统（PRES）**

该系统是一个包括盆地（凹陷）评价及局部圈闭评价的综合评价系统，这个系统的具体内容见后。

第一节 人工智能与专家系统

一、人工智能

人工智能（Artificial Intelligence，缩写为AI）是计算机科学中涉及设计智能计算机系统的一个分支，这种系统呈现出与人类的智能行为如理解语言、学习、推理和解决问题等有关的特性。按照这个概念，凡是使机器能够具有感知功能（如视、听、嗅）、思维功能（如分析、综合、计算、推理、联想、判断、规划、决策）、表达行为功能(如说、写、画）以及学习记忆功能等内容的都属于人工智能的研究范畴。

目前，将人工智能学科分为三个层次：

（1）人工智能的基础理论　凡是与人工智能有关的数学理论（如离散数学、模糊数学等）、思维科学理论（如认识心理学、逻辑与抽象思维学、形象与直觉思维学）和计算机工程技术（包括硬件和软件技术）都是人工智能的基础理论。

（2）人工智能的研究内容　知识的学习、获取、表达、处理，以及利用知识求解问题的基本技术都是人工智能的研究内容。

（3）人工智能的工程系统　根据人工智能原理而建立的具有实用价值的工程系统，都属于人工智能的工程系统。近年来，人工智能工程系统的主要研究课题有专家系统、自然语言理论、决策支持系统、模式识别系统、机器定理证明和机器人等。

现代人工智能技术的研究开始于1956年，首先是由美国的达特莫斯（Dartmouth）大学的麦卡锡（J. Mccarthy）与哈佛大学的明斯基（M. L. Minsky）等人共同发起的讨论机

器智能的会议，这次会议历时两个月，标志着人工智能这门新兴学科的正式诞生。

1969年成立了国际人工智能联合会议（JJCAI），许多国家也都成立了人工智能学术团体，例如美国的人工智能学会（AAAI），英国的AISB，意大利的GLIA等。1981年9月20日中国人工智能学会在长沙正式成立。

二、专家系统

专家系统（Expert System，缩写为ES）是人工智能的一个重要领域。美国斯坦福大学的费根堡（E. Feigenbaum）教授把专家系统定义为："一个使用知识和推理过程来解决那些需要杰出的专业人员才能解决问题的智能程序，它具备在这一层次上解决问题所必需的知识，加上推理过程，可以认为是对从事该领域的专家水平的模拟"。目前，专家系统的含义已经不仅仅等同于某个特定的专家了，而是应当理解为可以解决具有相当难度技术问题的"知识系统"。费根堡把建造专家系统的人称为"知识工程师"；把建造专家系统这一技术称为"知识工程"。

（1）专家系统的特征　专家系统与传统的程序系统有很大的区别，主要表现为：专家系统是一个计算机程序系统，主要由知识库和推理机组成，必须具有象专家那样解决实际问题的能力，例如咨询功能、学习功能、教育功能等。

（2）专家系统的应用类型　目前，专家系统可分为如下各种应用类型：解释型专家系统、诊断型专家系统、预测型专家系统、规划型专家系统、控制型专家系统、教学型专家系统等。

第二节　专家系统的基本结构

专家系统一般由五个部分组成，即：知识库(Knowledge Base)、数据库(Data Base)、推理机（Inference Engine）、解释部分以及知识获取部分。其中最重要的组成部分是知识库和推理机，见图3-5-1。

一、知　识　库

知识库是指存放在一起的，以某种形式表示的专家知识、经验以及常识的集合，以备系统推理判断之用。

知识是决定一个专家系统性能是否优越的主要因素。建立知识库决定于知识获取与知识表示这两个人工智能研究中的关键性问题。

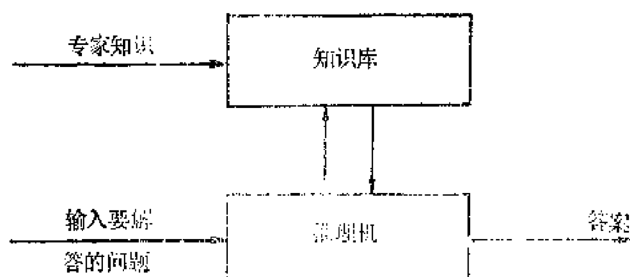


图3-5-1　专家系统的主要组成部分

二、数 据 库

数据库用于存储某个领域内初始证据和推理过程中所需要的基本信息，以及由系统得到的各种中间信息。例如在石油资源评价专家系统中，数据库存放有各种地质信息以及中间的评价结果等。

三、推 理 机

推理机是一组程序，是用来控制、协调整个系统，能够根据当前输入的数据，利用知识库中的知识，按一定的推理策略去解决当前的问题。推理方式有正向推理、反向推理以及正反向混合推理。在这三种推理方式中，又有精确推理与不精确推理，其中不精确推理是专家系统中必须认真对待的重要课题。

四、解 释 部 分

解释部分也是一组程序，可以对推理结果给出必要的解释，而使用户便于了解推理过程。这一功能充分体现了专家系统三大特征之一的透明性。

五、知识获取部分

知识获取部分是指为知识库扩充新的知识或修改原有知识所采用的技术手段。一般有三种方式，即：

（1）专家根据实际情况有所发现，而修改原有知识或增加新的知识。对于系统来说，这种方式的知识获取属于非自动方式。

（2）系统根据实践结果，能发现原有的某些知识有错误或需要增加新知识，并且通过专家去修改原有知识或增加新知识。这种方式的知识获取属于半自动方式。

（3）系统根据实践结果，自动地利用知识获取部分的功能，去修改原有知识或增加新知识。这种能够自动获取知识的方式，目前还很少能够实现。

第三节　知识表示方式

知识表示是人工智能各个应用领域和认识心理学所共同关心的基础问题。而知识表示方法是研究各种数据结构的设计，以便在知识库中存贮有关知识。

对于专家系统来说，所涉及的知识仅仅是现实世界中的一部分知识，这些知识包括：事物性知识、事件性知识、性能性知识和原知识。知识表示的实现，需要考虑表示能力、推理效率、正确性三个方面。表示能力是指能否将问题求解所需要的各类知识完全表示出来。推理效率是指能否有效地利用知识库中的知识进行完全推理。正确性是指知识表示方法是否具备准确定义的语义并保证推理的正确性。

目前，常用的知识表示模式有如下几种：

一、产生式系统

专家系统的知识表示方法中，产生式是最常用的一种。产生式系统由三个基本部分组

成，即数据库、产生式规则和控制策略，它们之间的关系如图3-5-2。

产生式规则是一个以“如果这个条件满足的话，就应该采取那个操作”形式表示的语句。凡是知识以孤立的事实和独立的规则集（合）表示的系统，称为纯粹的产生式系统，或称之为基于规则的系统。这类系统的特点是简单、表示形式单一、容易模块化、容易增添与修改；其缺点是效率低和具非透明性，非透明性是指人们很难跟随求解问题过程中的控制流。

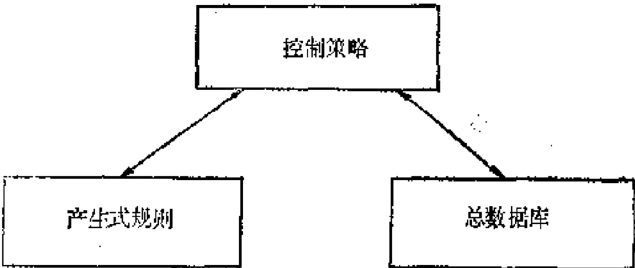


图3-5-2　产生式系统的组成

二、语义网络模式

语义网络最初是作为人类联想记忆的一个显式心理学模型。“语义”一词主要是指语言结构及其意义上的联系。后来，人们发现语义网络在逻辑推理方面的潜在能力，而使语义网络成为人工智能的一种知识表示方法，并且得到了迅速发展。

语义网络由节点和连接节点间的弧线所组成，其中节点表示实体、概念、情况等，而弧表示所连节点间的关系，而某些断言就可以用一个语义网络来实现。

三、过程表示方法

过程表示是把知识包含在若干个过程（小程序）之中。每个过程可以做到如何说明事物、事实以及其间的关系，在说明完好的情况下如何运行等方面的知识。这些程序不是以常规方法由其他程序调用，而是当某些条件成立时，由控制机构激活。

这种模式便于表示启发式知识，可以产生直接推理的特定论域信息。但是，相对于其他表示模式来说，过程模式实现的系统有其不完备性与不一致性，有时需要控制信息或牺牲知识库中的知识模块。

四、框架表示方法

知识的框架表示方法是由美国麻省理工学院的M. Minsky提出的一种知识表示的新理论，是将语义网、过程等知识表达思想结合起来，用于描述一些固定环境和常规行为。

Minsky从心理学的证据出发，认为人们在他们日常的认识活动中，使用了大量由以前的经验中获取并经过整理的知识，而这些知识是以一种类似框架的结构记存在人脑中的。框架提供了一种结构，对于新的数据人脑将用从过去经验中获取的概念来进行解释。并且，框架也是一种定型状态的数据结构，它的顶层是固定的，表示某个固定的概念、对象或事件；其下层是由一些称为槽（slot）的结构组成。通常框架表示为如下形式：

《框架名》

　　　　《槽名1》《侧面名11》（值111，值112，…）

　　　　　　　　《侧面名12》（值121，值122，…）

… … … … …

《槽名2》《侧面名21》（值211，值212，…）

《侧面名22》（值221，值222，…）

… … … … …

一个框架可以有任意有限个数目的槽，一个槽可以有任意有限个数目的侧面，一个侧面又可以有任意有限个数目的值。框架、槽、侧面可以描述各种各样的信息，而且侧面的值也可以是其他的框架。

第四节 不精确推理

不精确性在专家系统中是不可避免的，造成这种现象的原因主要有两点：一是推理依据的规则（或知识）不精确、不完善，而且对于不同学派来说也是不一致的；二是证据本身的不确定性。所以，在专家系统中往往要根据不充分的证据和不完全的知识进行推理。

对于处理不精确性问题，在60年代，多数专家系统是采用Bayes法，即以数理统计作为基本方法。由于这种方法需要大量的数据，并且不易给出先验概率而受到限制。因而自70年代开始，转向研究引起不精确性的原因以及解决这种不精确性的理论问题，并且提出如下一些理论方法。

一、主观Bayes理论

主观Bayes方法是由R. O. Duda等人于1976年提出的，是以概率理论为基础的一种方法，并于1978年应用于实际的专家系统，例如地质勘探专家系统PROCPECTOR就是以这种理论建立的专家系统。

这种方法的基本思想是：对断言H的信任程度应该随着新信息的获得而改变，也就是根据E的概率$p(E)$，利用规则（LS，LN），把断言H的先验概率$p(H)$更新为$p(H|E)$的过程。规则（LS，LN）中的LS是证据对假设的充分性度量，LN是证据对假设的必要性度量。这种实现从叶节点到假说的逐步推理过程，称为概率传播。

二、确定性理论

确定性理论是Shortliffe等人于1975年提出的，是以确定性理论为基础的一种方法，例如脑膜炎诊断专家系统MYCIN就是以这种理论建立的不精确推理系统。

三、证 据 理 论

1981年J. A. Barnett把由A. P. Dempster提出的并由G. Shafer改进的证据理论引入专家系统。这种理论是一种较老的方法，只需要满足比概率论更弱的公理系统，可以区分"不精确"和"不知道"这两种截然不同的情况。这种理论使用了可信函数，考虑了证据可能性的上、下界。有人认为证据理论的主要应用领域就是专家系统。

这种方法的不足之处在于证据间的独立性不容易保证，而且这种方法由于传递关系而引起的计算上的复杂性，使得在实现上比较困难。

四、可能性理论

可能性理论的基础就是查德提出的模糊集合理论。模糊集合可以认为是普通集合的一种扩充，它是由隶属函数来定义的。模糊数学与概率论之间有许多相似之处，概率论中的许多概念，在模糊数学中也应有其对应的概念，例如多重可能性分布、边界可能性分布、条件可能分布等等。

可能性理论的不足之处是存在着可能性之间的独立性问题以及先验可能性问题。

五、发生率理论

发生率理论是Bundy于1984年首先提出的，是用集合论中的元素内容来反应证据间的相关性。使用结合论结构，把命题间的逻辑运算变为结合运算。

这种方法的困难是要对集合赋值，而且发生率的计算只能给出下界值。因此，Bundy建议把证据理论的方法与发生率计算结合起来，提高推理效果。

六、假设推理理论

前面五种方法都是数值方法，然而，目前最活跃的是非数值方法，例如假设推理便是一种非数值方法，这种理论是由Doyle于1982年提出的。其中心思想是：不精确性是用列出规则的所有例外来消除的，如果不行，就用假设来描述默认值的特征及规则的可废除性，以便驳倒推理的能力。假设推理适于处理不完全信息，而对处理不精确信息则不太适用。这种推理主要有两种方法：其一是认为没有明显定义的断言，就认为其值为假；其二是当推出的某一结论与原来断言不矛盾时，就认为推出了这一结论。

这种理论的不足之处在于，它不能将概率信息与推理假设统一起来，同时结论也依赖于默认值的精确性。

七、定性推理理论

定性推理理论是由Cohen提出的，它是一种建立在保证条件下的纯粹定性理论。这里的保证是指基于命题证明的明确记录。保证可将证明加以分类，并且能为解释提供一个很好的机制，因为它们不仅产生而且保存了证明的整个过程。

八、证据空间理论

证据空间理论虽然目前还不很完善，但它有其独特性，它是把不确定性表示为二维空间中的点，两个坐标轴分别表示命题成立的可能性与不可能性。

第五节　PRES油气资源评价专家系统

PRES油气资源评价专家系统是由海洋石油勘探开发研究中心与吉林大学共同研制的，从1986年开始至1989年已基本建成了一个可以进行凹陷评价及圈闭评价的专家系统。它共由6个推理模块、9个知识库、3个数据库组成 。

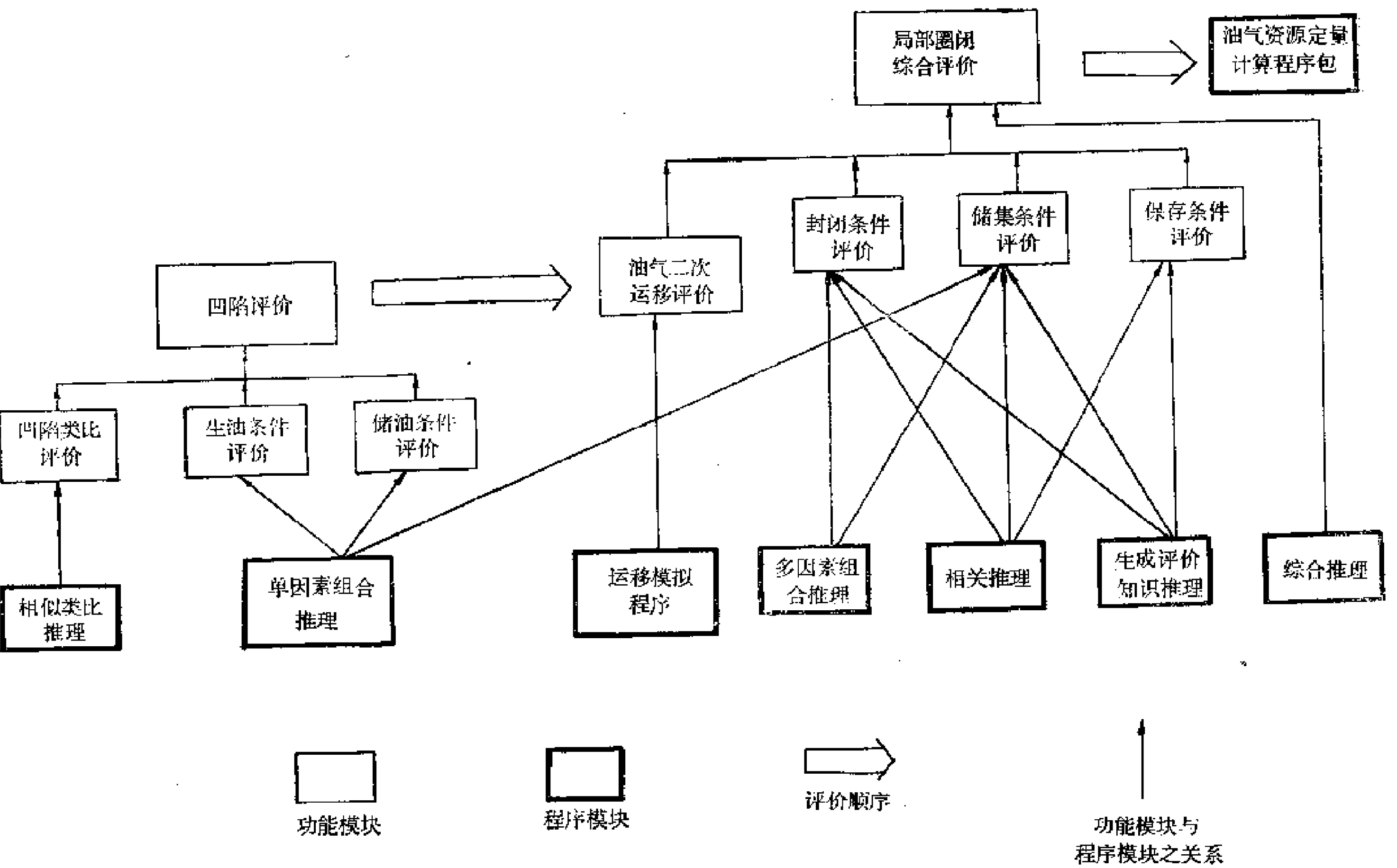


图3-5-3 PRES系统的原理与结构示意图

一、评价系统的内容及其要点

针对海洋油气资源评价的实际情况，PRES系统包括凹陷及圈闭评价两大部分。

凹陷评价部分包括如下子系统，其中凹陷类比评价子系统可以选择富油凹陷；生油条件评价子系统能够指出凹陷内的有利生油区，油气运移评价子系统是以生油评价与凹陷储集条件评价为基础，指出油气运移方向、时空配置关系、油气可能聚集的区带。而圈闭评价部分是完成每个圈闭的地质风险分析和油气资源量计算，见图3-5-3。

(1)为了适应在不同勘探阶段、不同地质条件以及不同资料情况下进行油气资源评价的需要，本系统是将多种常用方法以及各种可能获取的参数以任务分解、任务选择、任务排序和任务转换等知识表示方式灵活地加以运用。

(2)考虑到油气资源评价工作是在石油地质综合研究基础上，需要用多种学科知识进行分析、推理和判断的复杂工作，也就是说需要多个地质学家进行密切合作，因此PRES系统已在传统的专家系统基础上发展成为多知识表示、多推理机联合协作的第二代专家系统。

(3)在定性分析的全部过程中，地质参数的代表性、所用方法的适应性、专家知识的可靠程度等均有一定的不确定性。如何描述各个环节的不确定性，并且保证推理结论的正确性，正是专家系统不精确推理技术所要解决的关键性技术问题。为此，PRES系统采用了证据理论方法、多值可信度方法，较好地解决了上述问题。

(4)为使PRES系统的评价知识不断更新以保持其先进性，配备了知识获取和知识库维护功能，可以很方便地对知识库中所存贮的知识进行增加、删除或修改。

(5)油气资源评价工作在很大程度上依赖于基础资料的齐全、准确。PRES系统共有凹陷、局部圈闭、单井三个数据库，可以使系统自动读取其所需要输入的数据。

二、评价系统的原理简介

PRES系统共包括六个子系统，在此仅以生油条件评价、凹陷类比评价、油气运移与聚集条件评价为例，将系统的原理作一简单介绍。

1. 生油条件评价

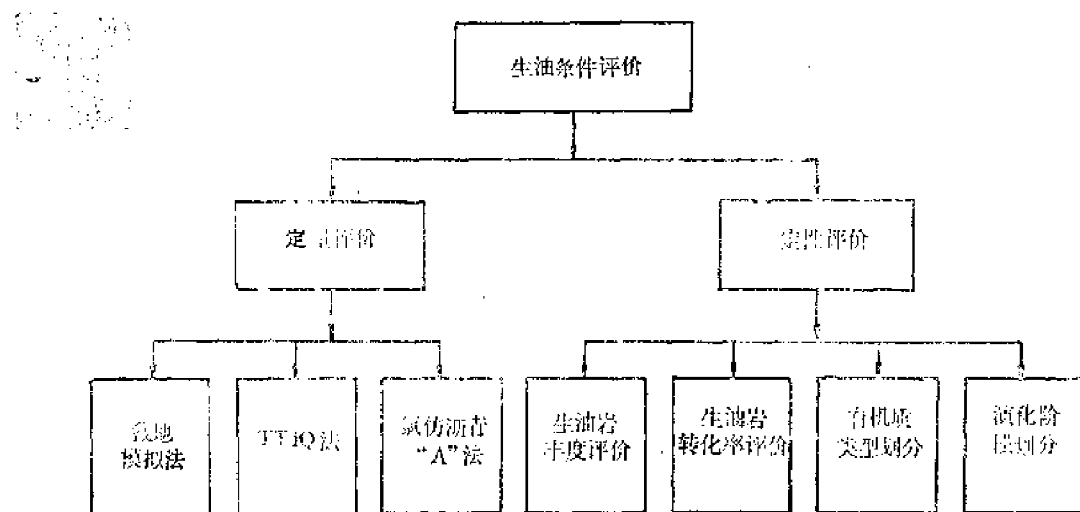


图3-5-4 生油条件评价任务分解图

凹陷内生油条件评价是作为该子系统待求解的总任务，这一总任务可以分解为定量评价和定性评价两个部分。定性评价还可以进一步分解为丰度评价、转化率评价、有机质类型划分以及演化阶段划分等子任务。而每个子任务还可以再分解为不同的求解方法。经过如此分解，则可形成一种树型结构，见图3-5-4。

这种树型结构只要通过修改分解规则，就可以任意添加或修改任务分解树中的节点；通过控制规则就可以根据不同地质情况来选取与组织求解方法；而通过转换规则可在反问推理时进行综合评价。

生油评价任务分解后的子任务，仍然可以细分为若干种具体任务。例如，有机质类型划分这一子任务（TORG），还可以分解为镜下鉴定法（TOMI）、热解法（TOSS）、沉积相法（TOSE）、元素分析法（TOAF）等，见图3-5-5。

图3-5-5中的叶结点是可以执行的一个方法。每个方法可以是以数值计算为主的方法；也可以是以定性分析为主的一组规则。下面列出的是用有机碳含量评价生油岩有机质丰度的一个方法，这个方法就是由一组规则所组成的。在PRES系统中，生油条件评价共包括有56个这样的方法，规则700余条。如：

〔方法名〕：用有机碳含量评价生油岩丰度

〔方法注释〕：作者：吴立真

时间：1985.8

〔对应任务〕：生油岩丰度评价

〔方法类型〕：以一般规则为主

〔适用条件〕：有机碳含量低于4%

〔衰减规则〕：

〔1〕如果：①有机碳含量大于0.6，且

②沉积相是沼泽相，或

③沉积相是河流相，或

④沉积相是滨湖相

则：可信度将衰减50%

〔2〕如果：①有机碳含量大于0.6，且

②有机质类型为Ⅲ型，或

③有机质类型为Ⅱ$_C$型

则：可信度将衰减50%

〔3〕如果：沉积相与有机质类型均无资料

则：可信度将衰减20%

〔操作规则〕：

〔1〕如果：①湖水咸度为淡水湖，或

②湖水咸度为半咸水湖，或

③没有湖水咸度此项资料

则：有机碳含量划分生油岩丰度的标准是：

有机碳大于1.0%应属于高丰度

大于0.6%，小于1.0%应属于较高丰度

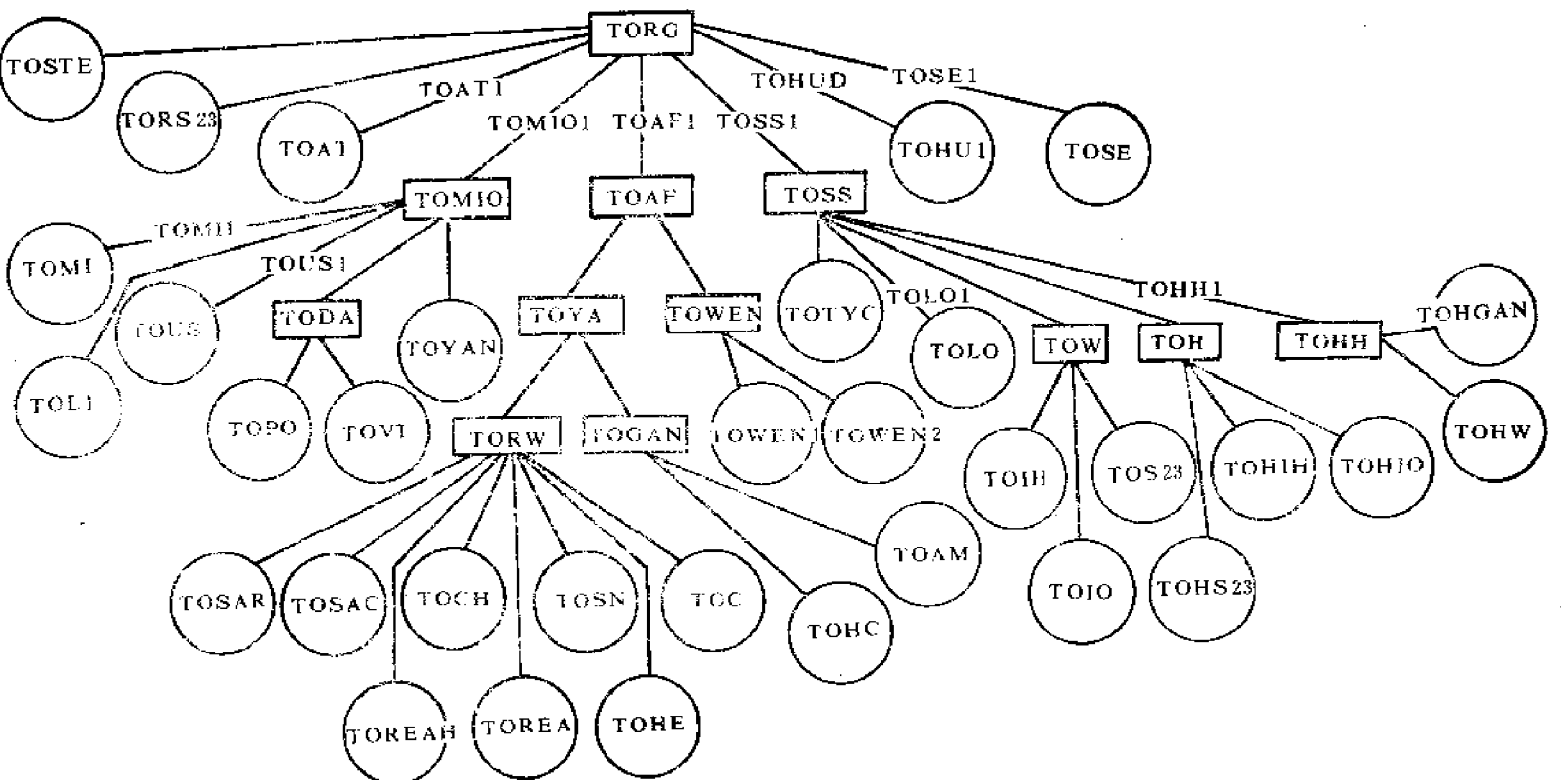


图3-5-5 有机质类型划分子任务的分解示意图

大于0.4%，小于0.6%应属于较低丰度

小于0.4%应属于非生油岩

〔2〕如果：①湖水咸度为咸水湖，或

②盐湖

则：有机碳划分生油岩的标准是：

有机碳大于0.6%应属于高丰度

大于0.4%，小于0.6%应属于较高丰度

大于0.2%，小于0.4%应属于低丰度

小于0.2%应属于非生油岩

〔一般规则〕：

〔1〕如果：有机碳含量大于1.0

则：有可信度〈C_1〉，证明应属于高丰度，同时

有可信度〈C_2〉，证明应属于较高丰度

〔2〕如果：有机碳位于〔0.6，1.0〕区间

则：有可信度〈C_1〉，证明应属于高丰度，同时

有可信度〈C_2〉，证明应属于较高丰度，同时

有可信度〈C_3〉，证明应属于低丰度

〔3〕如果：有机碳位于〔0.4，0.6〕区间

则：有可信度〈C_2〉，证明应属于较高丰度，同时

有可信度〈C_3〉，证明应属于较低丰度，同时

有可信度〈C_4〉，证明应属于非生油岩

〔4〕如果：有机碳小于0.4

则：有可信度〈C_3〉，证明应属于较低丰度，同时

有可信度〈C_4〉，证明应属于非生油岩

2. 凹陷类比评价

地质类比法是石油地质研究的一种基本方法，类比就是借助于与之相似的已知模型区去对未知评价区进行分析。

本模型主要是对凹陷进行类比，因此所选类比项主要是针对凹陷的地质结构、地层组成、剖面形态、沉积特征等有关方面。考虑到类比要适应不同的勘探程度，而采用了如果所能提供的类比项目内容较粗糙，就“粗比”，如果所能提供的类比项内容较细致，就“细比”。此外，每个项目的取值也有粗细之分，如果描述的较详细，类比的可信度就高，反之可信度就低。为此，将项目及取值都设计了一种树型结构。例如沉积岩厚度这个项目，当只能给出凹陷内沉积岩的一个总体厚度概念时，可以用此项直接参加类比。如果在剖面中能分出沉积的早、中、晚三个阶段，就可以分别用三个阶段的厚度去进行对比，这显然会增加类比的可信度。如果再能分出每一阶段的平均厚度或最大厚度进行对比，就会更大地增加类比的可信度。见图3-5-6。

对于项目的取值也是采用树型结构。例如凹陷的类型为断陷时，再细分可能取值为单断或双断，双断又可再分出对称和不对称双断，而对称或不对称双断又可细分为中凹、中隆、中斜等。不论是哪一级的取值都可以进行类比，只是对比结果的可信度不同，见图3-5-7。

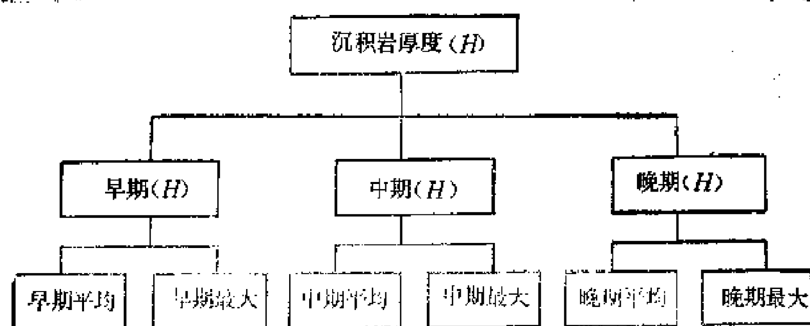


图3-5-6 沉积岩厚度的树型结构图

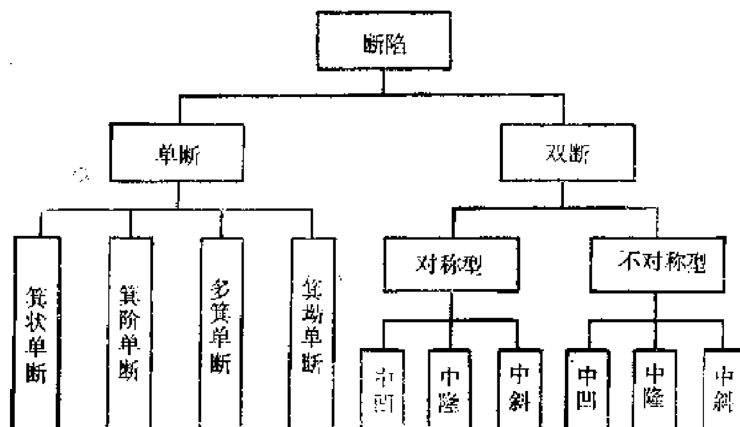


图3-5-7 断陷的树型结构图

3. 油气运移与聚集条件评价

本系统是根据地质专家对油气运移的各种认识，利用所能得到的地质参数，从宏观角度来分析油气运移的通道、运移的方向、时空配置关系，从而得到评价区内所有圈闭油源条件的相对好坏。系统所采用的控制因素及输入输出内容列于表3-5-1中。

表3-5-1 油气运移与聚集条件评价表

控制油气运移的主要地质因素	输入资料	输出结果
凹陷（盆地）内生、排烃潜力及空间分布	各节点在不同历史时期、不同层位的排烃量	评价区任意比例尺的构造发育史图
浮力	评价区内各层等厚图	任意方向横切的构造发育剖面图
输导体	各层系砂泥岩百分比等值图或沉积相图	不同历史时间不同层位的油气运移方向图
断层	断层的平面位置及现今剖面图	不同历史时间不同层位的油气运移规模图
由压实水流产生的水动力	若干已知井的孔隙度与深度关系图	各个圈闭的运聚规模与构造发育的配置关系图
		各个圈闭的运聚规模与构造发育参数表

该子系统所采用的方法是将评价区分层位划分出适当间距的网格单元，每个网格单元作为运移评价的基本单元。

划分运移单元之后，对于某个单元内油气运移量的模拟及运移方向的判断可按下述规则进行：

（1）评价区内可同时存在多个生油中心，本系统可在统一的网格体系中，按统一的地层层序处理。

（2）每个单元都可能有其自身的排烃量，从其他单元运移来的烃类数量，向其他单元运移出去的烃类数量以及在本单元内散失的烃类数量，这些量均可按着下面的（5）、（6）、（7）规则计算。

（3）从所有单元中顶面高度最低的单元开始起算，将该单元运移出来的量存入运移指向的单元。依照从低到高的顺序直至所有单元均计算一次。

（4）每个单元的运移以构造的低部位向高部位为指向，但是同时要考虑指向单元是否能成为运移通道，如遇断层则按不同情况分别处理；如遇相带变化则用相应规则判断是否改变运移方向。

（5）每个单元所流入的烃类数量可能包括压实水动力形成的量、横向或垂向的运移量、通过断层的输导量以及沿不整合面运移来的量。每个量是以控制其相对大小为主。

（6）每个单元所流出的烃类数量与规则（5）包括的内容相同。

（7）假设在不形成油气聚集条件下，所有单元均有一定的散失量，则散失量的大小决定于单元内的储集空间。

（8）将不同地质时期的圈闭作为油气聚集的主要场所。圈闭包含的所有单元均假设先充满其储集空间，如有多余的量才作为流出量，因此，每个圈闭都要确定溢出点。对于断层或岩性形成的遮挡条件，本系统暂不考虑。

（9）系统运行一次只是对某个地质时期已经沉积了的所有地层。如果进行某个圈闭的油源条件评价，可将所有地质时期流入该圈闭的量相加，就得到所有可能流经该圈闭的量。

在实现上述规则时，将一些控制参数作为可变参数，例如：控制运移方向的砂泥比界限，控制散失量大小的系数，压实水动力所形成分量的比例等。这些可变参数可通过已知区的模拟来确定。由于圈闭的油源评价结果是个相对量值，因而不必强求可变参数的精确性。

该系统曾在珠江口盆地试用，评价目标是惠州凹陷及其邻区。当时，将整个地层分为文昌、恩平、珠海、珠江组下段、珠江组上段、韩江、第四系共7个层系。经过系统自动恢复古构造和油气运移方向的模拟，最后预测了22个已知圈闭的运移聚集量。将这些预测量与圈闭的地质储量相比较，其中的17个基本吻合，这说明该系统的预测结果是有参考价值的。

三、PRES专家系统的软件结构

将专家知识与推理程序进行分离是专家系统与一般计算机程序的主要区别，并且这种结构展示了越来越广泛的应用前景。如何使推理软件具有通用的工具性，以便于开发更多的实用性专家系统，已成当今世界的研究重点。本系统就是一个用专家系统工具自动生成的实用系统，其中MES1就是本系统内的工具软件之一。

在研制PRES系统过程中，研制者不仅力图用一种推理去完成全部所要达到的评价目的，而且要着眼于资源评价的各个方面，并考虑整个石油地质领域的知识特点，将知识划分

出不同的类型，每种类型的知识采用相应的知识表示方式以及推理形式。也就是说，一种知识表示及推理形式既可以解决一个评价目标，同时又可以解决相同类型的多个评价目标，这样的推理软件显然具有工具性。其中相似类比推理是一对一的关系，而单因素组合推理则是一对多的关系。具体来说，就是用单因素组合推理加生油条件评价的知识库可以解决生油条件评价；加储集层条件评价知识库可以解决储集层条件评价。而其他软件模块也与此相似。

按此种结构来讲，PRES系统共包括了6个目标级专家系统。PRES系统的总体结构可以表示为多用户界面环境、第二代协作专家系统以及辅助系统三个部分，见图3-5-8。

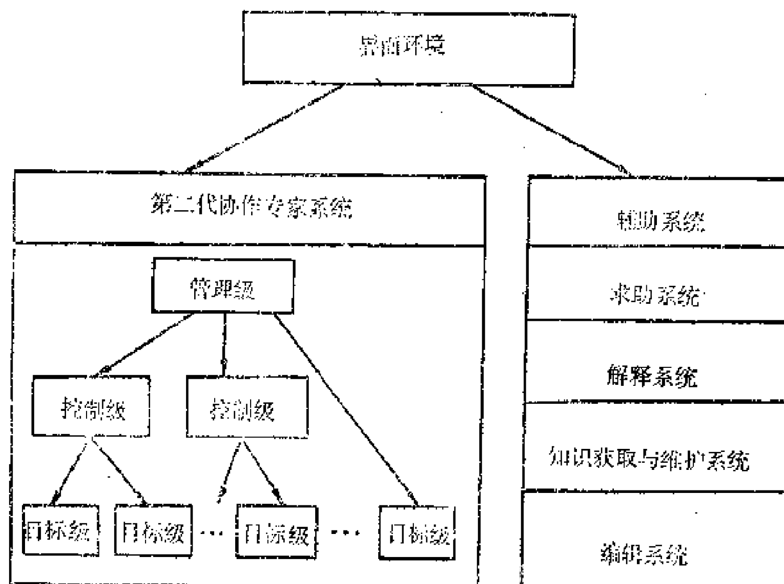


图3-5-8　PRES系统的软件结构示意图

目标级专家系统是具有独立推理机，并且连结相应的知识库和数据库而构成的完整系统。表3-5-2中列出了其推理机的功能及其与知识库、数据库的关系。

下面以生油条件评价系统为例，说明其软件的工作原理。生油条件评价系统是在MES1工具系统支持下建造的，图3-5-9表示了这种支撑关系。

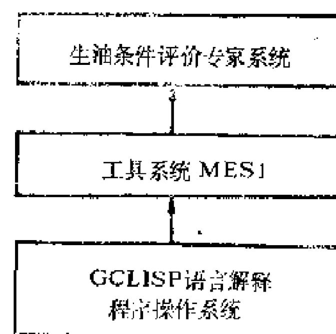


图3-5-9　生油条件评价系统与工具软件MES1之间的关系示意图

MES1工具软件与其他工具系统相比，具有如下特点：

（1）从系统结构上看，由MES1所支持建造的专家系统具有元级和目标级的两级结构。对于用户提出的任务首先由元级系统经过推理，确定目标系统的动作，然后由目标级系统具体执行而得出评价结果。

（2）元知识库与元推理机分离。在元级系统中，是将元知识与元推理机分离，形成元知识库。元知识主要由任务分解规则、任务选择规则、任务排序规则、任务转换规则等6种元规则组成，这些元知识主要是用来描述与控制知识。由于元知识与元推理机分离，使得系统具有较好的灵活性和较有效的控制结构。

（3）对于陈述性、操作性知识的表示是以方法为单位、每个方法中包括可信度衰减规则、操作规则、过程规则，控制规则等13个条目。每个条目可以由若干条知识组成。

表3-5-2 PRES系统各模块的功能表

推理机程序模块			知识库模块		数据库模块	
名称	字节(k)	功能	名称	功能	名称	字节(k)
单因素组合(MES1)	300	使用多个有关的单因素或方法以解决评价目标的定性分类问题	生油凹陷定性评价	划分生油凹陷的类型	凹陷库	
			单井(点)生油条件评价	进行凹陷内生油条件定性评价	单井库	290
			凹陷储集层定性评价	划分凹陷储集层类型	凹陷库	
			单井(点)储集条件评价	进行凹陷内储集条件定性评价	单井库	425
			圈闭储集条件评价	进行局部圈闭的储集条件评价	圈闭库	
多因素组合		用互相有制约关系的诸因素共同确定评价目标的性质	圈闭封闭条件评价	进行局部圈闭封闭条件评价	圈闭库	
			圈闭储集条件评价	进行局部圈闭储集条件评价		
相似类比(GAES)	250	按相似原理将未知体与若干已知体类比以加深对未知体的认识	生油凹陷类比	通过与已知生油凹陷类比求得未知凹陷生油量	凹陷库	
相关推理		按因素间的相关特征推断事物的未知属性	评价因素相关关系	用以解决缺少参数时的评价		
油气运移	65	模拟油气二次运移的方向及运聚规模				
资源量计算程序包	220	计算盆地凹陷和局部圈闭的资源量				

(4)关于不确定性推理，在MES1系统中采用了A. P. Dempster与G. Shafer提出的证据理论关于不确定性的处理方法，并且进行了扩充。

(5)MES1系统除了对知识库中的知识可以进行插入、建立、编辑、删除、修改以外，还使用了基于元知识的辅助知识获取，见图3-5-10。

(6)对推理过程、系统行为等给出解释是专家系统的重要特征之一，也是区别于传统程序的一个重要方面。生油条件评价系统中的解释功能主要包括询问项目取值、解释概念及定义、说明评价方法、给出评价结果、显示推理过程等等。

第六节 对专家系统的展望

目前，美国、日本以及我国的有关单位都在积极研制新一代的专家系统。新一代的专家

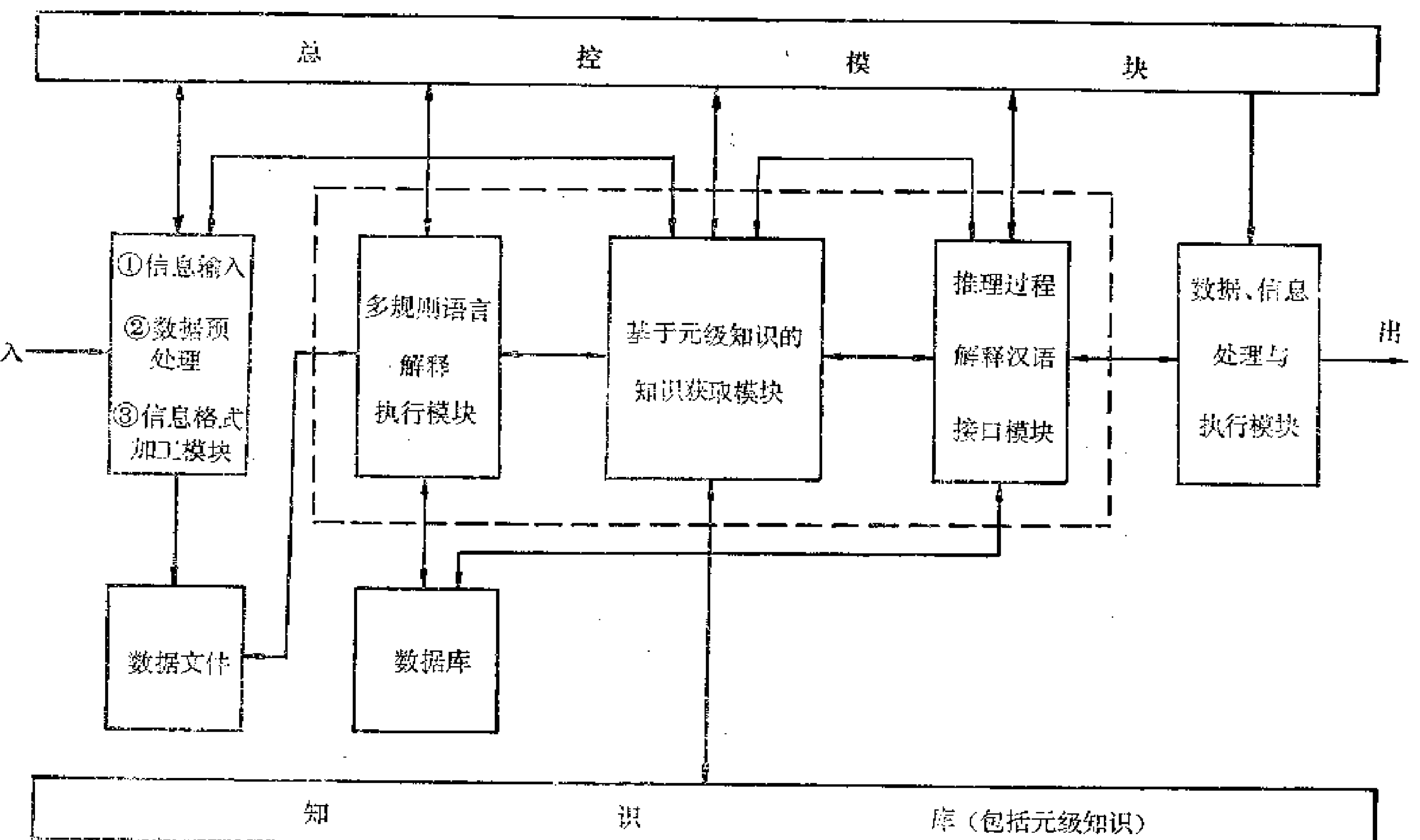


图3-5-10 生油评价专家系统结构图

系统其处理知识的能力在质和量两个方面都远远优于第一代专家系统。主要的差别是新一代的专家系统采用并行技术，使用了新的专家系统专用语言，配备智能知识获取子系统，提供建造专家系统的工具系统，具有智能人机接口，包括专业汉语和专业英语的理解。

从系统的种类方面来看，可能向以下几个方向发展。

一、大型综合系统

这种系统将能组合各种知识表达和推理的策略，而理想的工具则可允许用户模拟所有的咨询模式，例如推理既可正向链接，也可以逆向链接，而知识表达既可用框架表示方式，也可以用产生式规则，以使各类知识库可以协同工作。

二、大型专门系统

大型专门系统是指用只有一、二种特殊咨询模式的工具所开发的专家系统。EMYCIN是典型的专门工具，它是一种允许用户以逆向链构造产生式规则的系统。这类工具可在某一特定模式上构造复杂的编程环境供系统开发使用。预计今后一段时间内将会涌现很多种专用系统。

三、专用工作站

专用工作站是指那种可以帮助管理人员和专业人员完成大量工作的个人计算机或智能化工作站。这种工作站的实现将取决于各种小型知识系统是否能够组合在一起工作。一个专用工作站应该是一个综合系统。

四、小型专门系统

小型系统是指具有100条到500条规则的系统，一般可在个人计算机上运行，主要用途是帮助用户处理小规模的疑难问题。预计近期这类系统可能会得到广泛普及，其中最可能的用途是指导用户解决一些不太熟悉、但又具有某些规则可循的工作问题。

五、固化的专家系统

目前已有许多单位在研究专家系统的硬件固化问题，固化后形成智能化仪表，例如电脉冲解释系统就可以通过一个芯片来解决。今后这方面的应用可能会越来越重要。

第六章　石油地质数据库

数据库技术始于本世纪六十年代初期，1963年美国HONEYWELL公司开发的IDS—Ⅱ软件产品是世界上公认的最早的数据库。该软件的主要设计者巴克曼（Bachman）曾指出：数据处理以程序为中心是不合理的，应改为以数据为中心，以集成数据库为中心。自此建立了数据库的概念，确立了这项新技术的重要地位。此后，数据库技术一直是计算机科学中非常活跃的一个领域，与其他计算技术的发展相辅相成。

70年代初，美国IBM公司的E.F.Codd发表了题为"大型共享数据库的关系模型"的文章，提出了关系模型的概念。他在总结网状数据库的基础上使关系式数据库得到了完善，使之逐步成为当代数据库的主流模型。ORACLE、DATABASE2（DB—2）、DBASE（Ⅱ、Ⅲ、Ⅳ）、INGRES、INFORMIX、RBASE等20余种关系式数据库管理系统已成为流行软件，而且版本不断更新，功能不断增强。

三十年来，数据库技术经历了60年代的文件系统，70年代的多模型数据库系统，80年代的关系式数据管理系统三个发展阶段，目前已成为一门比较成熟的技术。当今世界已有数以万计的数据库系统在运行，其中有些已联结成为国际数据通讯网络。按目前数据库的发展趋势，今后数据库技术将侧重于分布式数据库系统、数据库计算机、知识库、公用数据库结构、数据库逻辑设计自动化等方面。

为了提高地质资料的使用效率和管理水平，促进地质研究工作的定量化，数据库技术已在地质界得到普遍重视，并已作为地质学现代化的一个重要标志。60年代中期，欧美一些国家开始建立地质数据库；80年代以来，随着计算机技术的飞速发展，使得地质数据库技术大面积普及应用。我国的地质数据库技术虽然起步较晚，但在最近十年已有很大进展，目前各主要油田均已建立了规模不同的地质勘探数据库，存储的内容以钻井、试油资料为主。国内的大型地质数据库也正在积极筹建之中。

用户对于数据库的基本要求是数据的存取效率高、运算速度快。因而，研究数据的组织方案，便成为数据库技术的重要课题。一个数据库的设计过程可分为两个阶段，即逻辑设计阶段和物理设计阶段。逻辑设计是指根据数据库的设计理论和用户特定的专业要求，研究确定数据的结构、层次和相互关系。物理设计是指从机器的角度出发，设计出数据在存储介质上的最优存取途径。本章将重点讨论石油地质数据的逻辑设计。

第一节　数据库的逻辑设计

以石油地质数据库为例，逻辑设计阶段的任务是从油气勘探开发工作流程和专业数据的特点出发，研究数据的结构、层次和相互关系，提交数据的最佳组织方案和用户对数据完整的描述和定义。

一、石油地质数据库逻辑设计的步骤和内容

（1）从石油地质的专业工作出发，研究地质数据的用户概念模型，并且根据数据库的设计原则对地质数据进行新的逻辑分类，推导出数据组织的树型结构图。

（2）依据数据结构图确定数据库的各个文件，建立文件间的相互关系和关系间的联络手段，进一步修改或优化全局逻辑结构，导出各级用户层的逻辑模式。

（3）从完善数据库的操作、管理和应用出发，对全部数据项进行标准化设计。逻辑设计阶段实施的标准化主要是针对字段信息的准确完整描述。

（4）逻辑设计应为数据库总体设计提交下述文件：全局逻辑结构模式图、全局文件结构模式图、数据库文件结构设计表、数据项标准化设计表（逻辑设计词典）、数据库服务指示器设计表等。

（5）数据库的常规应用设计也应当归入逻辑设计部分。它包括常规应用的基本框图、运算统计公式、基础表格和图件的格式和要求等。常规应用设计既是数据组织结构设计的参照系，也是对逻辑设计质量检验和修正的标准。

逻辑设计是数据库设计的首要环节和基础工作。它不仅与数据词典设计、数据库物理设计、应用设计、数据资源的分布式管理设计密切相关，而且要直接指导庞大、复杂的数据资料整理入库工作。

数据库设计工作必须按数据定义所要求的设计目标来进行。目前，数据库的定义是：一组存储在一起，具有最小冗余度的，互相联系的数据。能以最优方式为尽可能多的应用服务，数据的储存应与具体使用数据的程序无关，要使用一个公共方案可增添、修改、删除、检索库内的已有数据。

二、数据库的设计目标

（1）最小冗余度，即要求消除储存数据中的有害重复以及可能的隐含数据。

（2）数据的独立性，即数据项的任何逻辑和物理结构的变化，均不会影响对库内数据的调用。

（3）数据项之间的联系性，即数据项之间应定义它们的从属关系、互访关系、运算关系、顺序关系，对于关系模型又要求每一个信息描述段都表示关系的一个单一概念（即范式化）。

（4）数据组织的结构化，并且具有扩充性。

（5）数据组织是多目的多元统一体，即数据应能为多用户多用途服务，数据组织要有专业上的完整性和应用上的统一性。

（6）对数据的操作要灵活、快速、安全、可靠、简单、易行。

对于每个具体的数据库系统，可能其所面临的专业问题不同，使用的计算机设备不同，采用的数据库管理软件不同，数据库的设计方法不同，但是，在实现上述六个目标方面应该是一致的。

第二节　石油地质数据库设计实例

本节以全国石油天然气勘探开发数据库一个实验性设计方案为例，介绍数据库的设计方法。

一、地质数据信息的概念模型

数据库的逻辑设计是从回答下面两个问题开始的：

（1）数据库的用户（或持有者）是如何把客观体系概念化的？

（2）用户的实际问题是什么？

这两个问题的实质是要求把数据库描述的事物恢复为原始形式，以便从信息的角度讨论数据的结构问题。对石油地质数据库而言，就是要从调查分析石油地质数据的范畴、来源、格式、现行分类法、资料间的关系和联系、资料的运行流程以及其他生成数据资料的控制机制入手，得出石油地质信息的概念模型和对石油地质信息的描述规律，以便找到一种逻辑关系清晰的、结构化的、概念化的石油地质信息分类规则。

石油地质勘探中获取的多种信息，具有范围广、内容多、数量大的特点。常用的石油地质数据可分为8个方面约40类。8个方面是：构造地质、钻井地质、测井、试油测试、地层、分析化验、采油、综合。这种专业分类方法也是地质界长期延续下来的收集、整理以及使用地质数据的传统方法。但是，从数据库的定义来看，这种分类有许多缺点：其一是为保持每一类数据的内容完整，许多数据项之间将会因为相交重复而造成较高的冗余度，而且不同资料内相同项目的数据内容仍会不一致；其二是从地质概念和资料系统化方面分析，传统分类其数据之间的联系与线索也比较模糊。

鉴于上述情况，在传统资料分类的基础上应提出一个简化的模型，见图3-6-1。这个可包含常用石油地质数据内容的简化模型称为实体模型，此处所说的实体是指信息所描述的对

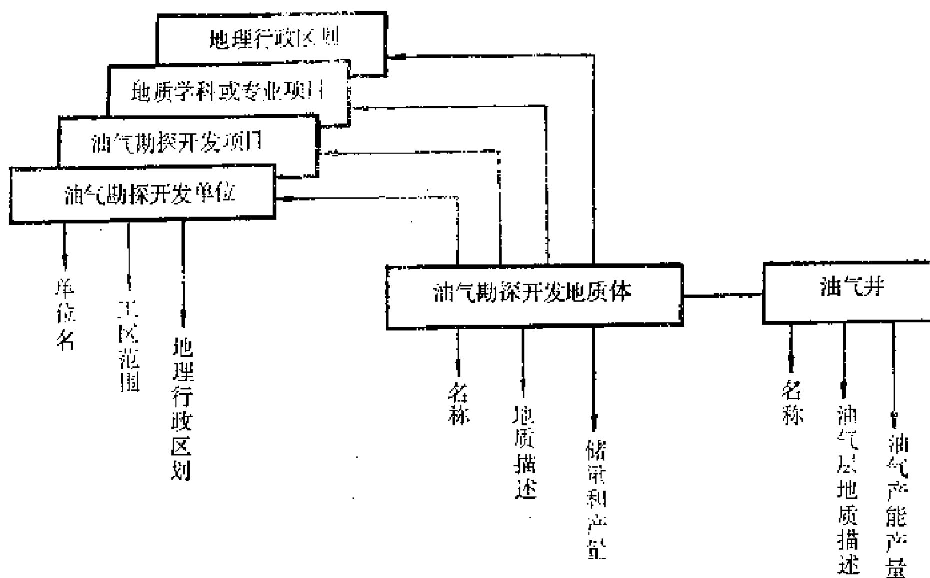


图3-6-1　实体模型

象。

从实体模型出发可以得到石油地质数据的一个新定义：石油地质数据是用来描述和研究各种地质体及其油气形成、产出规律，并在地质体的三维空间中找出它们之间内在联系的，所采集的可为多用户服务的一组多元（或多项目）信息。

用这一定义来分析石油地质数据之间的联系与线索，可以把全部数据项分为四大类。

1. 地质体的命名数据

如沉积盆地、坳陷、凹陷、圈闭、断层、油气田的名称；储集层、油气层的地层层系名称；井号、试油层号、分析化验样品编号等。

2. 地质体的描述数据

如圈闭要素、岩性、油气产状、沉积特征、地层时代、地层温度压力、地球化学分析、岩石物性、古生物、岩矿鉴定等。

3. 产出物的描述数据

如试油、采油过程中产出物的名称、产出量、产出时的温度压力及其变化、储量、产出物的分析化验数据等。

4. 辅助性数据

辅助性数据是指对地质体或产出物进行观察、测试、分析时，有关条件、方法、过程以及工作量等方面的非地质数据。例如时间、井身结构、管柱结构、试采工作制度、取样条件、施工状况等。

上述这种分类法可以使地质数据之间有较为清晰的逻辑关系，这是导出地质数据树型结构的基础。

二、石油地质数据的树型结构和数据库的逻辑模式

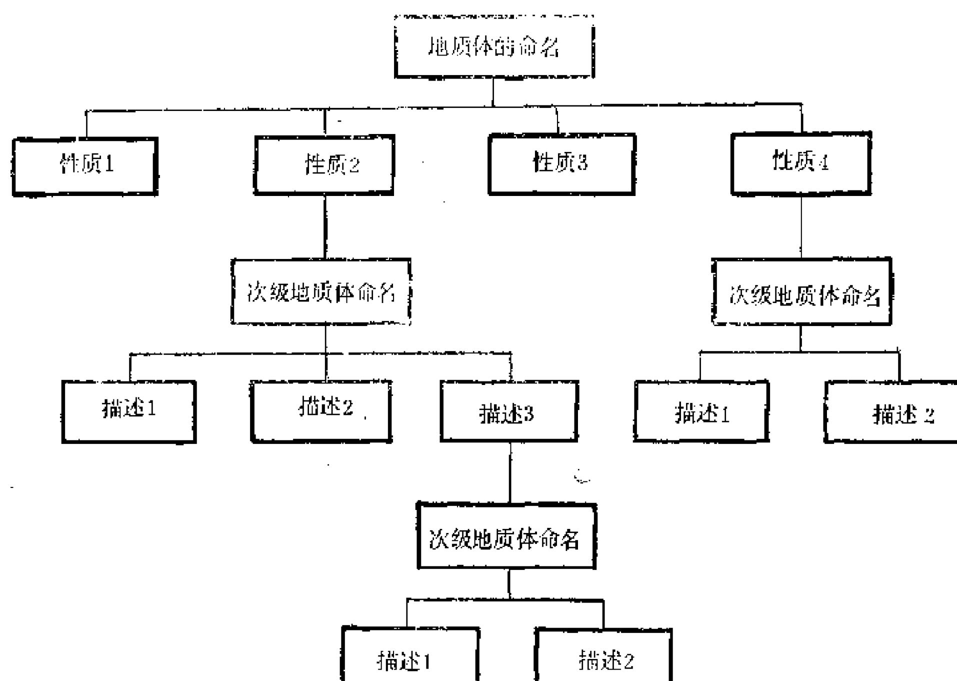


图3-6-2　地质数据的简化树型结构模型

在上述地质数据实体属性的分类基础上，可以导出下面的简化树型结构模型，见图3-6-2。

树型结构是数据库较普遍的一种结构形式。它的根和每个结点就是系统中的各个数据文件，而根和其中的部分中途结点将是地质体的命名文件或其他主要关键字文件。

在这种树型结构基础上，可以进一步设计数据库的各个子系统模式。每个子系统应该是一组性质类似的数据集合。除了各种以地质体命名的子系统外，还应考虑在勘探开发过程中起管理或监控作用的一些子系统。石油地质数据库可由如下7个子系统组成。

1．勘探开发的行政单位子系统

勘探开发的行政单位子系统包括的内容为承担勘探开发任务的各级行政单位名称及其隶属关系，见图3-6-3。

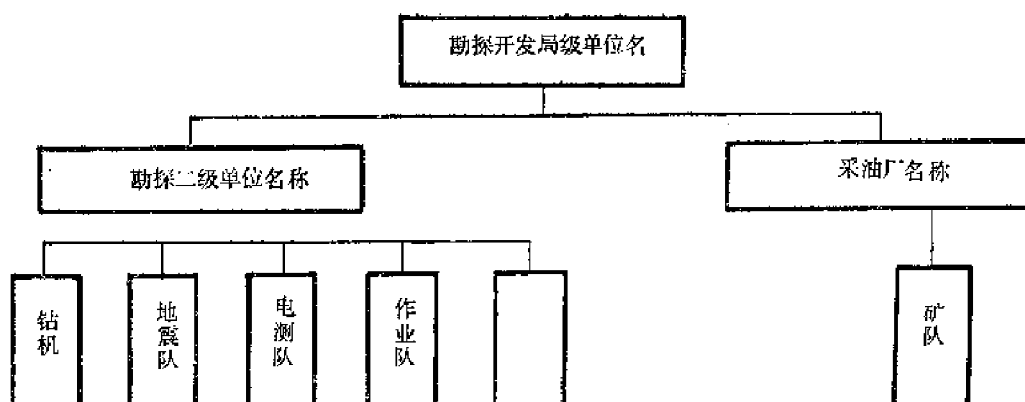


图3-6-3　勘探开发的行政单位子系统框图

2．勘探开发年鉴子系统

勘探开发年鉴子系统所包括的内容为：勘探开发部署、产量任务、施工记录、新增产量及储量、大事纪要等，见图3-6-4。

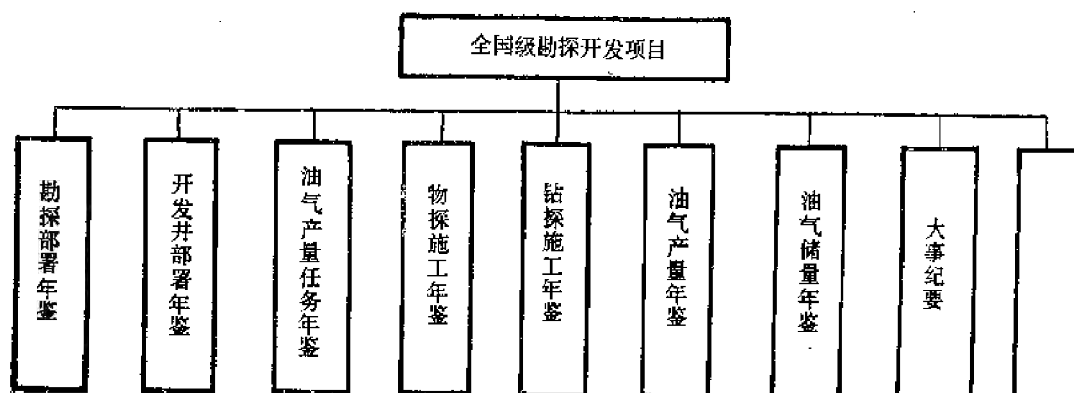


图3-6-4　勘探开发年鉴子系统框图

3．构造描述子系统

构造描述子系统所包括的内容为：构造单元名称、各级构造单元的基本要素及其相应的评价结果等，见图3-6-5

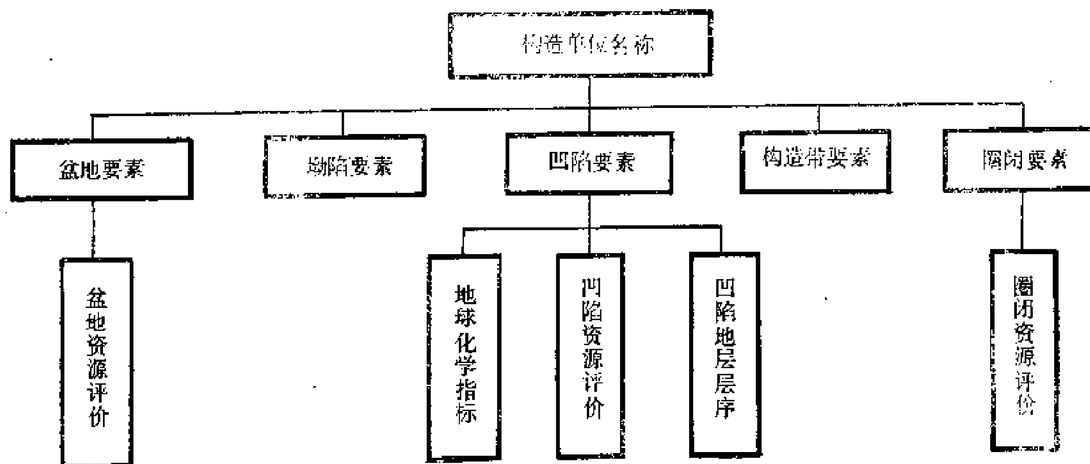


图3-6-5 构造描述子系统框图

4. 油气田描述子系统

油气田描述子系统所包括的内容为：油气田的基本数据、油气藏参数、油气储量、产量记录、油气层层序、区块划分等，见图3-6-6

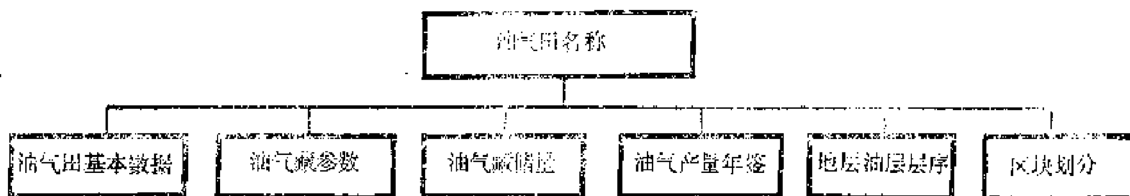


图3-6-6 油气田描述子系统框图

5. 钻井地质描述子系统

钻井地质描述子系统包括的内容较多，有：钻井基本数据、录井油气显示、电测解释结果、试油结果、岩石化验结果等，见图3-6-7。

6. 注采井描述子系统

注采井描述子系统所包括的内容有：某个具体注采井的油气产量记录、注水量记录以及与注采层段有关的增产措施记录等，见图3-6-8。

7. 油气矿产登记子系统

油气矿产登记子系统所包括的内容有：区域勘查、工业性勘探、滚动勘探开发、油气开采等登记表，见图3-6-9。

以上简要地介绍了每个子系统的主要内容，这些子系统中的每个子系统大多数都可以成为大型分布式数据库中某个结点工作站的数据系统模式。例如，钻井描述子系统就是一个钻探实体的基本数据模式；注采井描述子系统就是采油厂矿数据系统的模式。如果用实体命名指示器和系统服务指示器并适当地联结各个子系统，便可以构成油气勘探开发地质数据库的逻辑模式图。在未进行文件设计之前，它只是一个简化图，反映的内容是用户易于理解的数据组织轮廓以及数据描述的各个实体之间的从属、指向关系。这种图中可有百余个结点，也就是

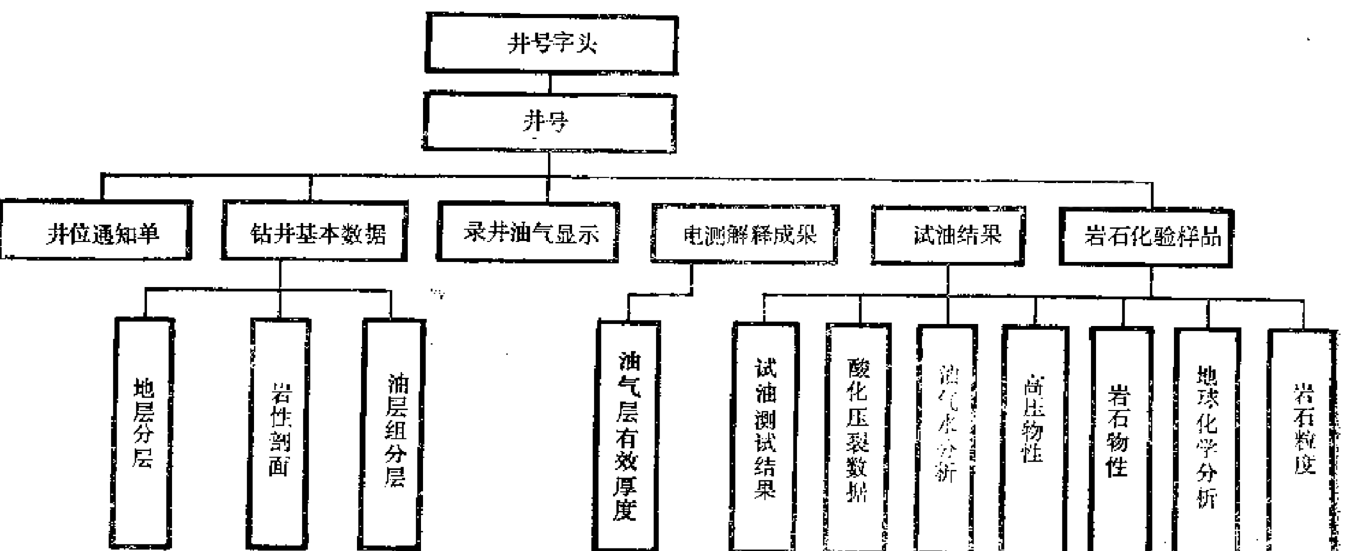


图3-6-7　钻井地质描述子系统框图

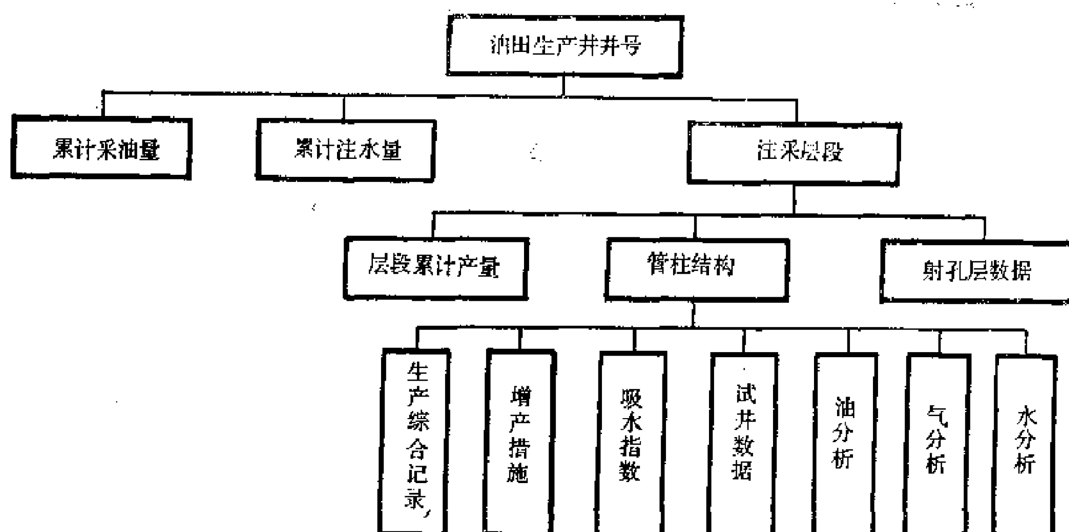


图3-6-8 注采井描述子系统框图

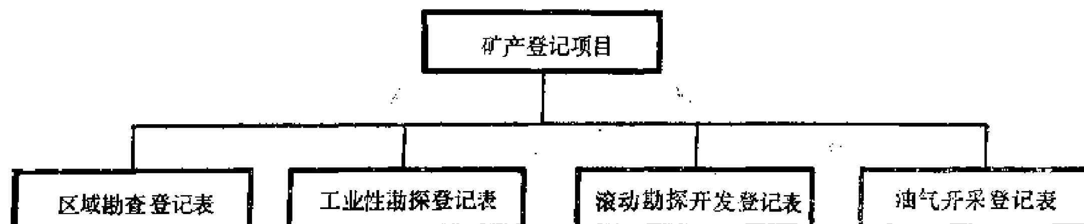


图3-6-9 油气矿产登记子系统框图

说，地质数据库也将有同样数目的文件组织。这些文件组织逐一确定后再纳入模式图，就构成数据库完整的逻辑模式图。这种图可以作为数据库设计的导航图。通过分析逻辑模式图，认为地质数据库的总体结构有如下特点：

（1）一个完整的石油地质数据库的结构是一个具有多级多叉的非平衡树，有百余个结点。

（2）树型结构的二级结点命名代表了地质数据的基本分类或者地质数据的集合方式，总共有三种形式：①企业名称（或持有者名称）；②勘探开发管理或施工项目；③地质体或矿产体。地质数据库的最简单结构框架见图3-6-10。

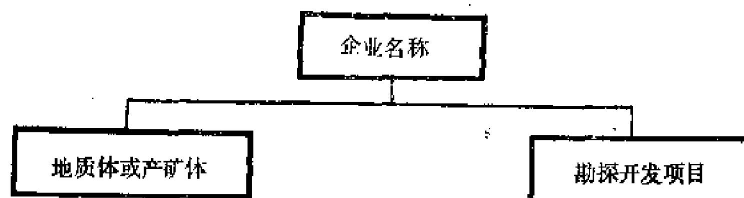


图3-6-10 地质数据库的最简结构框架图

（3）由上面的图3-6-10可以导出数据系统的骨干框架，见图3-6-11。正确的骨干框架可使逻辑模式返回实体模型的形式，这便于检验逻辑结构的合理性和完整性，也就可以对数据系统进行更深入的分析。

（4）在逻辑结构图中，说明数据子系统之间联系的导航是由指示器文件完成的。指示器是专门负责数据库索引、查询的一种文件组织形式，原为物理设计中的一个概念，后被逻辑设计所采纳，而成为一种特殊的逻辑型文件。指示器的类型和功能有如下三种：

①主关键字指示器文件　这种指示器文件由主关键字名称、骨干属性、子系统联络子段、

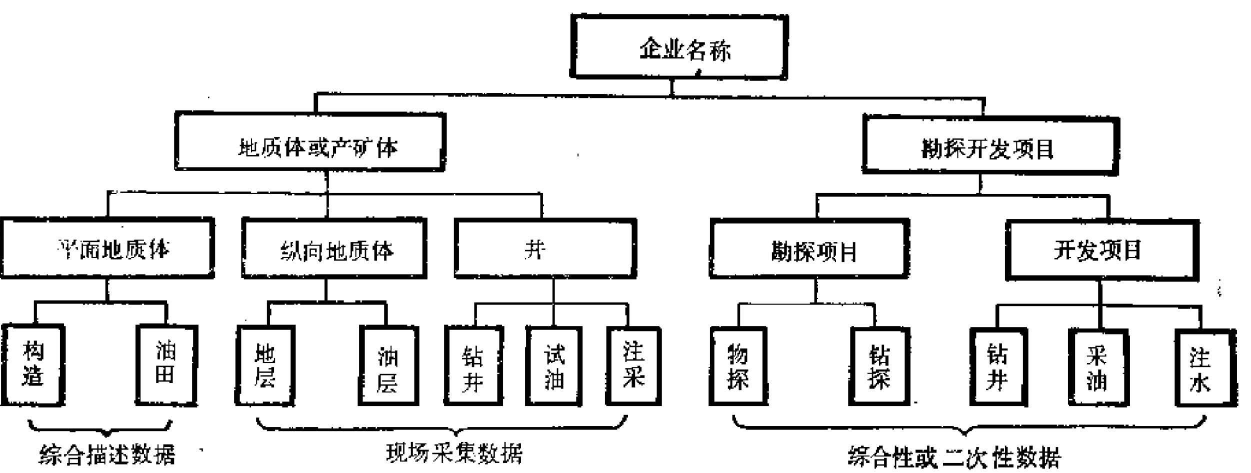


图3-6-11　数据库系统的骨干框架

服务性数据、查询地址等内容组成，其中数据查询子系统的切换功能等应具有灵活性和快速反应性。

②标准命名指示器文件　这种文件是为实施数据系统的标准化而编制的，具有服务和查询等功能。例如：地震测线字头指示器、井号字头指示器、地层代码指示器等。这种指示器也可用于对地区性数据文件结构的说明，例如地层分层文件的字段名与相应地区的地层代码指示器的连接。

标准命名指示器的查询功能见图3-6-12。

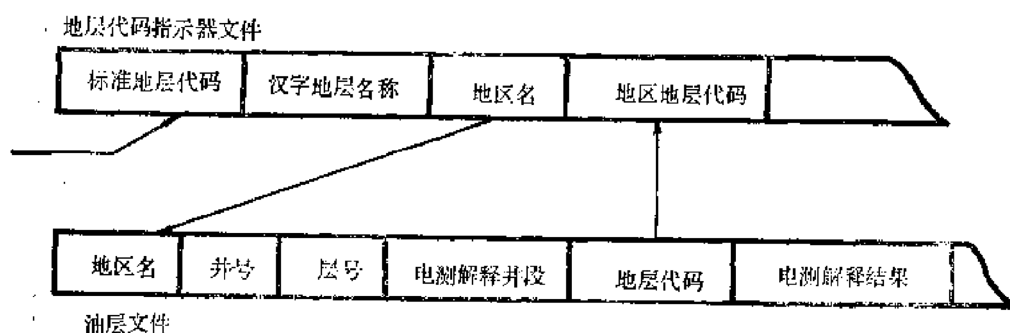


图3-6-12　查询功能示意图

标准命名指示器对文件结构的说明功能见图3-6-13。

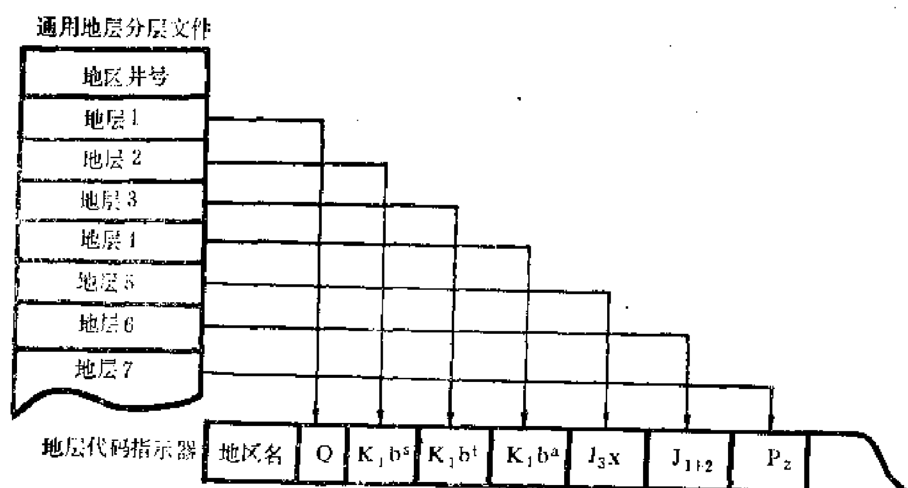


图3-6-13　标准地层命名指示器对文件结构的说明功能示意图

③服务性指示器　这种文件的功能是数据输改、运算、成果输出以及其他管理等。例如，井别指示器的功能有输出汉字井别、印制图形、确定顺序号与运算编号等，见图3-6-14。

指示器文件在数据库中身兼多能，具有消除冗余度、改变数据结构等多种功能。例如，在构造名指示器中建立各级构造单元归属字段后，便把构造要素的多级结构变成同级多维的形式。

数据库全局逻辑结构模式图是数据库设计的基础图件，它将在研制数据库的管理和应用软件时被反复使用，并在设计和建库过程中不断地得到修改和优化。

输入标识符	输出汉字	印制图形	顺序号	运算编号
QT	区域探井		1	1
YT	预探井		2	2
PJ	评价井		3	3
SC	生产井		[illegible]	4
CY	采油井		5	4.1
ZS	注水井		6	4.2
GC	观察井		7	4.3

图3-6-14　井别指示器功能图

三、数据文件结构和文件模型的范式化

数据文件又称数据集或记录集，一个数据库就是若干数据文件的有机集合。数据文件是由相同格式的数据段（记录）所组成，见图3-6-15。文件设计的第一步是根据逻辑结构框架逐一将各结点与命名有关的资料内容展开，由若干字段组合成该文件的数据段；第二步是分析数据段的逻辑结构特点。各字段间的归属传递关系和填入数据的有关规定，并根据关系模型进一步修改逻辑结构，使数据段的设计尽可能地规范化。

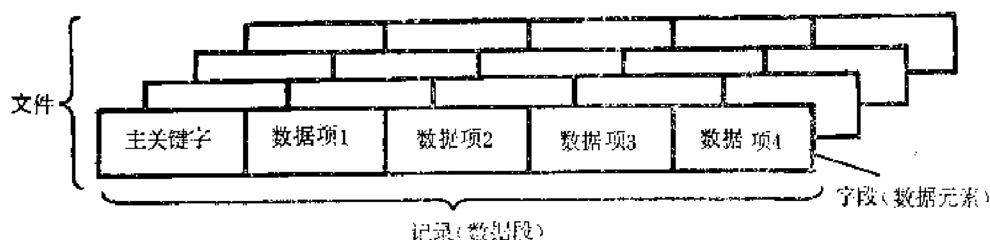


图3-6-15　数据文件的结构图

数据段的设计，实际上就是依据逻辑模式将各结点上有关的地质资料项目展开划段，划段时应参考下列7条标准：

（1）把实体命名及其有关的骨干数据项和描述实体性质的数据项分开划段。

（2）把有业务交接功能的项与无交接功能的项分开划段。

（3）把使用频度和更新频度较为一致的划为一段。

（4）把静态数据与动态数据隔开划段。

（5）把同一实体内重复出现与不重复出现的数据项分开划段。

（6）使每段内的数据项排序合理。

（7）划段和排序要尽量与原始资料的书面归档格式相似，以利于数据的录入。

数据段是一个等待填入数据的框架。在同一描述对象的控制下，数值在各种数据段框架

中的循环方式和出现形式不同，我们称之为段型，地质数据段的段型有4种。

1．单一定长型

同一描述对象只有一个数据项相同且相等的数据段则为单一定长型，例如钻井基本数据段，见图3-6-16。

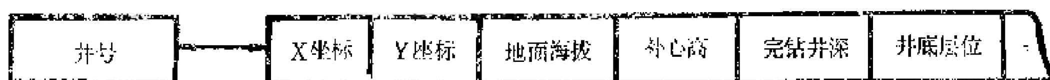


图3-6-16　钻井基本数据段

2．多重定长型

此种类型表征的是同一描述对象有若干个数据项相同且相等的数据段，但是，不同对象的数据段个数可能是不同的，例如钻井取心数据段，见图3-6-17。

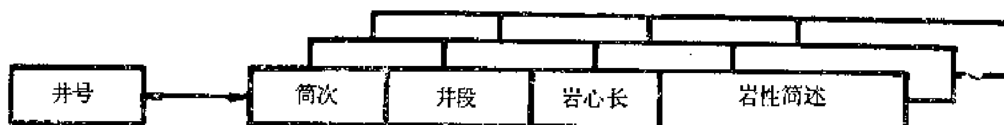


图3-6-17　钻井取心数据段

3．单一变长型

此种类型表征的是每个描述对象只有一个数据段，但这个数据段中数据项的名称和个数是可变化的，如套管程序段，见图3-6-18。

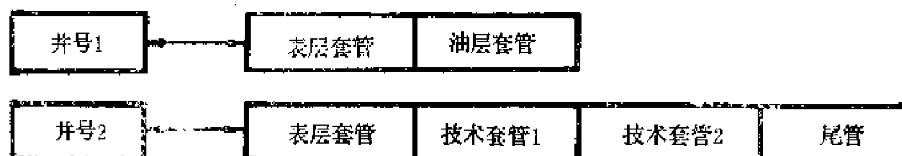


图3-6-18　套管程序段

4．多重变长型

表征的是每个描述对象都有若干个数据段，段数和各段的数据项个数都是变化的，如钻井油气层综合解释数据段，见图3-6-19。

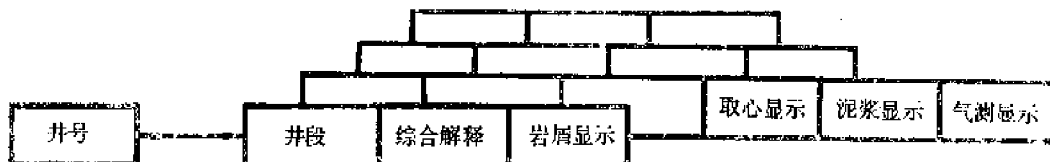


图3-6-19　钻井油气层综合解释数据段

当然，段型是可以变化的。通过改变描述对象、组段方式或改变数据项的定名、类型，后三种段型都可以变成单一定长型，即二维表格的平面类型。讨论数据段型的目的是要把非范式的文件结构向范式化转换，而产生出更合理的数据文件模型，提高存储效率，确保数据结构的稳定性，使其易于操作。

数据结构的范式化设计方法应来自实现数据库定义目标的长期实践，它主要应解决如下两个突出的技术问题：

（1）提高数据库管理软件对数据管理、操作和运算的自动化程度，要求数据结构易于用程序语言描述，管理程序有实现数据导航的功能。

（2）计算机网络和数据库系统的结合，使数据库的存取途径由集中式转为分布式，这就需要一种公用的数据结构对各分布结点的数据资源进行统一管理。

从巨大工程是由简单工程的组合而成的原理出发，设计人员应当从错综复杂的数据结构中，回到"平面文件"这种最普通和最简单的文件形式上来。平面文件就是把一个数据关联于一个数据项，把这个数据项关联于一个固定的属性，再把一组相关的属性关联于一个实体，数据组织的每一行与一个实体有关，每一列与一个属性有关。平面文件的特点是没有重复组与重复行，各列命名相异，而每列类型相同。用平面文件组成的数据库就是关系式数据库。关系式数据库的逻辑设计一般采用E-R法（Entity-Relation），即实体关系设计法，这种设计法的步骤与本例基本近似。

在关系模型推导过程的最后，还有结构范式化设计。此处的范式化是指数据组织平面结构的最优形式。目前，通常将范式化分为五级标准，一般认为设计达到第三范式便可以满足数据系统的要求，更高的范式有可能使系统出现其他缺陷和不足。关系模型数据文件范式化的过程如图3-6-20所示。该范式化的本质是把前面分析过的四种数据段型都转变为单一定长型，常用的转换方法有如下四种：

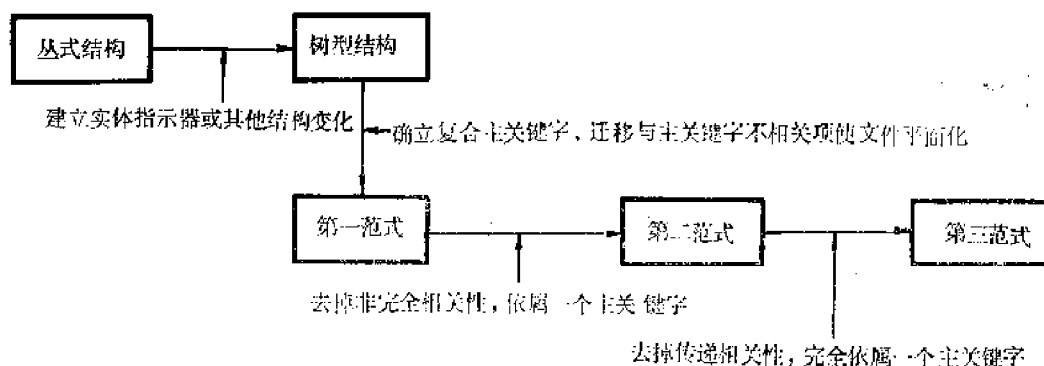


图3-6-20　关系模型数据文件范式化流程图

（1）改变分段方式，把复杂的表格分解化简，使之成为平面表格。

（2）改变描述对象，把平面地质体命名和纵向层系命名组合成主关键字，如井层、井样品等。

（3）改变数据名称和类型，增加备注字段，消除重复组。

（4）按施工期的阶段建立多期文件，即前期文件采用低范式结构，后期文件用高范式结构，前期文件完成后用程序自动生成后期文件。

任何范式的变化都是对数据组织进行地质逻辑分析的过程，离开对数据的专业化分析，就难以建立合理、清晰的关系模型。图3-6-21是一个文件范式化设计的实例。

油气勘探开发数据库文件设计均可经过范式化变成平面文件结构。但是，对每种文件转化为何种范式为最佳，则应按用户层的差异，计算机系统的差异，数据库管理软件的差异等讨论决定。对较大型的地质数据库的总体逻辑设计，应推出向下兼容度较高的文件组织方案，以

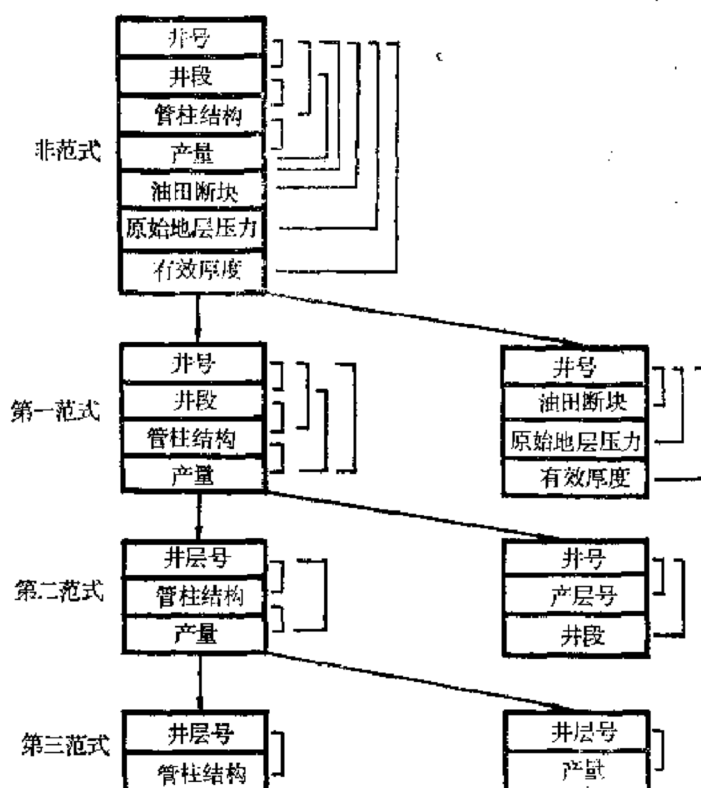


图3-6-21 范式化设计实例

利于数据库系统末端结点的建设。

四、数据库的字典设计

数据文件确立后，要在每个文件的数据字段定名的基础上，对每个字段进行定义和说明。要对数据项的地质含义、数据特征、值域与采集、审定方法等进行准确的说明。数据字段的定义和说明包括下列项目。

（1）数据字段编号，即字段在文件中的顺序号。

（2）文件标识名，即字段所属文件的数据库系统引用名。

（3）字段标识名，即数据库系统引用该字段的标识名称。

（4）汉字段名。

（5）字段类型，取数值型、字符型、日期型、逻辑型中的任一种。

（6）字段长度。

（7）小数位数。

（8）实体说明，即该字段是对哪个实体进行描述。

（9）属性说明，即该字型的地质含义或说明。

（10）填值说明，数值型字段要说明其值域范围；字符型字段则要说明应用类型是文字、标识符、图形或印刷体中的哪一种。

（11）单位说明，若同字段填值单位不统一，则应说明各种单位的类型、识别标志和互相换算的表达式。

（12）取值条件。某些字段的值有取值条件，如原油性质中的粘度值，应说明条件的表示方法。

（13）审定标准，说明字段的填值内容以何种资料的哪一项为审核标准，以便统一资料的出处。

（14）数据操作权，对数据进行输入修改等工作的一级操作用户名称。

第三节　地质数据库设计的后期工程

数据库逻辑设计是一个对设计方案反复修改、多期完善的过程。一个全国性的石油天然气勘探开发数据库的设计工程应分四期实施。

一、第一方案设计

要在勘探开发的各个实施单位征求修改意见，并同时进行数据库实体命名文件（构造、油田、地层命名、井号字头、地震测线字头）的信息普查。

二、第二方案设计

要建立系统的全部实体命名文件。选定若干数据库技术基础好的探区和油田，按新方案改造完善各自的数据库系统，校验和考核逻辑设计。

三、第三方案设计

建立全国一级数据库工作站的骨干信息系统，开展远程数据通讯网和分布式数据库资源共享试验，进一步论证方案的可行性，并着手设计数据库管理和常规应用软件。

四、最终方案设计

数据库系统工程全面实施，其中数据库管理和常规应用软件的设计是后期工程的重点之一。这些软件应该是一个符合油气勘探开发地质工作流程的，具有自动化办公特色的数据接收、修改、管理以及常规性、日常性资料处理输出、信息咨询等多种功能的应用软件。各级数据库在初级建设阶段，在很大程度上要依赖这个软件系统。它应能较全面地体现逻辑设计的标准和目的。

参 考 文 献

〔1〕刘承祚、孙惠文编著，《数学地质基本方法及应用》，地质出版社，1982年。

〔2〕王学仁编著，《地质数据的多变量统计分析》，科学出版社，1982年。

〔3〕方开泰、潘恩沛著，《聚类分析》，地质出版社，1982年。

〔4〕朱裕生，《矿产资源评价方法学导论》，地质出版社，1984年。

〔5〕余金生、李裕伟著，《地质因子分析》，地质出版社，1985年。

〔6〕翁文波著，《预测论基础》，石油工业出版社，1984年。

〔7〕油气资源评价方法研究与应用编委会编，《油气资源评价方法研究与应用》，石油工业出版社，1988年。

〔8〕于志钧、赵旭东编著，高等学校教学用书，《石油数学地质》，石油工业出版社，1986年。

〔9〕赵旭东，《石油资源定量评价》，地质出版社，1988年。

〔10〕徐钟济编著，《蒙特卡罗方法》，上海科学出版社，1985年。

〔11〕汪培庄编，《模糊集合论及其应用》，上海科学技术出版社，1983年。

〔12〕M.费史著，王福保译，《概率论及数理统计》，上海科学技术出版社，1962年。

〔13〕数学手册编写组，《数学手册》，人民教育出版社，1981年第2次印刷。

〔14〕赵旭东等著，《中国数学地质》(1)，地质出版社，1986年。

〔15〕赵旭东，"用Weng旋回模型对生命总量有限体系的预测"，科学通报，第32卷第18期，1987年。

〔16〕赵旭东、张守本，"应用数学地质方法对二连盆地进行石油资源评价"，石油学报，第6卷第3期，1985年。

〔17〕赵旭东、黄旭楠，"勘探方案的线性规划模型"，石油学报，第9卷第4期，1988年。

〔18〕傅京孙，《人工智能及其应用》，清华大学出版社，1987年。

〔19〕林尧瑞，《专家系统原理与实践》，清华大学出版社，1988年。

〔20〕J.阿尔迪，《专家系统》，上海科学技术文献出版社，1988年。

〔21〕程惟宁，"数据库管理系统软件的发展与现状"，数据库通讯，1984年1期。

〔22〕史忠植，"国外数据库技术概况"，数据库通讯，1983年2期。

〔23〕赵晨等摘译，"关系式数据库评测"，计算机世界报，1989年9月13日。

〔24〕James Martin，"计算机数据库组织"，计算技术通讯，1979年1-2期。

〔25〕Iv.Flores，"数据结构和管理"，计算机与图书馆，1979年2期。

附　　录

这里收录了用FOR TRAN-77语言编写的源程序共14个，包括本书第一篇第三章地质数据预处理的有关算法，第二篇地质多元统计分析的各种算法。限于本书的篇幅，第三篇的算法程序未能收入。为便于读者使用，作如下说明：

(1) 在每个源程序的开头，对程序的功能、程序符号、数据文件、输出文件都作了较为详细说明。在源程序中必要之处作了汉字注解，分段说明编程含义。

(2) 这些程序是选用电子工业部第六研究所推出的汉字DOS作为操作系统，用汉字WS编辑的，用FOR1（3.3版本）、PAS2、LINK（3.02版本）进行编译可形成执行程序，可在IBM-PC/XT、AT主机与宽行打印机构成的最低硬件配置条件下运行。

(3) 可执行程序在汉字DOS操作系统下运行时，显示器上有中文提示指导用户操作，可以通过汉字驱动程序由宽行打印机输出汉字计算结果。

(4) 程序之后附有用来检测数据验算的计算结果。为印刷之便，对计算结果的编排作了适当的调整，如字符图的列数均比源程序的实际输出列数少了一半。

程序一　变量标准化方法

一、程序主要功能

本程序包括对变量进行标准化变换的七种方法，即：总和标准化，最大值标准化，模标准化，中心标准化，标准差标准化，极差标准化，极差正规化。

二、程序符号说明

N———————样品数（要求N小于500）；
M　——————变量数（要求M小于20）；
X(500,20)——原始数据矩阵，矩阵的行号为样品编号，矩阵的列号为变量编号；
Y(500,20)——变量经过标准化变换后的数据矩阵；
X1(500)———计算平均值，模，总和的工作单元；
X2(500)———计算极小值的工作单元；
X3(500)———计算极大值的工作单元；
X4(500)———计算标准差的工作单元；
KK —————标准化的选择方式；
　　KK=1———————总和标准化；
　　KK=2———————最大值标准化；
　　KK=3———————模标准化；
　　KK=4———————中心标准化；
　　KK=5———————标准差标准化；
　　KK=6———————极差标准化；
　　KK=7———————极差正规化。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为Z1.DAT，如果使用其它名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
----------------------------------
N,M
((X(I,J),J=1,M),I=1,N)
----------------------------------
```

例如，下面的数据文件(Z1.DAT)就是一个供用户检测本程序的数据文件：

```
----------------------------------
5,4
1000.,500.,1.5,2000.,
250.,150.,1.0,2200.,
100.,70.,3.0,1500.,
10.,200.,2.0,1800.,
40.,100.,5.0,2500.
----------------------------------
```

四、计算结果输出

本程序输出文件的约定名为Z1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、变量标准化方法的主程序

```
      PROGRAM Z9001
      COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
      CHARACTER FILENA*20,LL*1,KK*1,NOYES
      INTEGER KK1
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'变量的标准化'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名Z1.DAT]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z1.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名 [约定名Z1.WRI] ：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z1.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入原始数据矩阵：'
      READ(1,*,ERR=5)N,M
```

```
      READ(1,*,ERR=5)((X(I,J),J=1,M),I=1,N)
      WRITE(*,'(26X,A)')'您对变量进行何种标准化?'
      WRITE(*,'(10X,A)')
     -'------------------------------------------------------------'
      WRITE(*,'(10X,A)')
     -'1,总和标准化    2,最大值标准化    3,模标准化    4,中心标准化'
      WRITE(*,'(10X,A)')
     -'5,标准差标准化  6,极差标准化      7,极差正规化'
      WRITE(*,'(10X,A)')
     -'------------------------------------------------------------'
    2 WRITE(*,'(10X,A\)')'请您选择:[1--7]=?'
      READ(*,'(A)')KK
      IF(KK.LT.'1'.OR.KK.GT.'7')GO TO 2
      KK1=ICHAR(KK)-48
      WRITE(*,*)'开始进行变量的标准化计算;请等待!'
      WRITE(2,*)'          * * * * * * * * * * * * * * * * * * * * * *
      WRITE(2,*)'          *                                           *'
      WRITE(2,*)'          *           变量的标准化计算结果            *'
      WRITE(2,*)'          *                                           *'
      WRITE(2,*)'          * * * * * * * * * * * * * * * * * * * * * *'
      WRITE(2,100)N,M,KK
  100 FORMAT(//5X,'样品数=',I3/5X,'变量数=',I3/5X,'变量的标准化方式=',
     -A3)
      WRITE(2,'(//11X,A)')'原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      DO 3 I=1,N
    3 WRITE(2,102)I,(X(I,J),J=1,M)
  102 FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      CALL XN0(N,M,KK1)
      WRITE(2,'(//11X,A)')'变换后的数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
      DO 4 I=1,N
    4 WRITE(2,102)I,(Y(I,J),J=1,M)
      CLOSE(1)
      CLOSE(2)
```

```
      WRITE(*,'(1X,A\)')'程序运行完闭：还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    5 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

(1)变量标准化方法的子程序

```
      SUBROUTINE XN0(N,M,KK1)
      COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
      INTEGER KK1
C     根据KK1确定变量的标准化方式.
      GO TO(1,2,3,4,5,6,7)KK1
C     总和标准化
    1 CALL XSI0(N,M)
        DO 10 J=1,M
        DO 10 I=1,N
   10   Y(I,J)=X(I,J)/X1(J)
      RETURN
C     最大值标准化：
    2 CALL XMAX0(N,M)
        DO 20 J=1,M
        DO 20 I=1,N
   20   Y(I,J)=X(I,J)/X3(J)
      RETURN
C     模标准化：
    3 CALL XMO0(N,M)
        DO 30 J=1,M
        DO 30 I=1,N
   30   Y(I,J)=X(I,J)/X1(J)
      RETURN
C     中心标准化：
    4 CALL XCP0(N,M)
        DO 40 J=1,M
        DO 40 I=1,N
   40   Y(I,J)=X(I,J)-X1(J)
      RETURN
C     标准差标准化：
```

```
 5 CALL XCP0(N,M)
   CALL SF0(N,M)
   DO 50 J=1,M
   DO 50 I=1,N
50 Y(I,J)=(X(I,J)-X1(J))/X4(J)
   RETURN
C  极差标准化:
 6 CALL XCP0(N,M)
   CALL XMAX0(N,M)
   CALL XMIN0(N,M)
   DO 60 J=1,M
   DO 60 I=1,N
60 Y(I,J)=(X(I,J)-X1(J))/(X3(J)-X2(J))
   RETURN
C  极差正规化:
 7 CALL XMAX0(N,M)
   CALL XMIN0(N,M)
   DO 70 J=1,M
   DO 70 I=1,N
70 Y(I,J)=(X(I,J)-X2(J))/(X3(J)-X2(J))
   END
```

(2)计算总和的子程序

```
  SUBROUTINE XSI0(N,M)
  COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
  DO 1 J=1,M
  X1(J)=0.
  DO 1 I=1,N
1 X1(J)=X1(J)+X(I,J)
  END
```

(3)计算模的子程序

```
  SUBROUTINE XMO0(N,M)
  COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500), X4(500)
  DO 2 J=1,M
  X1(J)=0.
  DO 1 I=1,N
1 X1(J)=X1(J)+X(I,J)*X(I,J)
2 X1(J)=SQRT(X1(J))
```

```
      END
```

(4)计算平均值的子程序

```
      SUBROUTINE XCP0(N,M)
      COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
      DO 2 J=1,M
      X1(J)=0.
      DO 1 I=1,N
    1 X1(J)=X1(J)+X(I,J)
    2 X1(J)=X1(J)/N
      END
```

(5)计算标准差的子程序

```
      SUBROUTINE SF0(N,M)
      COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
      DO 2 J=1,M
      X4(J)=0.
      DO 1 I=1,N
    1 X4(J)=X4(J)+(X(I,J)-X1(J))**2
    2 X4(J)=SQRT(X4(J)/N)
      END
```

(6)计算极小值的子程序

```
      SUBROUTINE XMIN0(N,M)
      COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
      DO 1 J=1,M
      X2(J)=X(1,J)
      DO 1 I=1,N
      IF(X(I,J).LT.X2(J))X2(J)=X(I,J)
    1 CONTINUE
      END
```

(7)计算极大值的子程序

```
      SUBROUTINE XMAX0(N,M)
      COMMON X(500,20),Y(500,20),X1(500),X2(500),X3(500),X4(500)
      DO 1 J=1,M
      X3(J)=X(1,J)
      DO 1 I=1,N
      IF(X(I,J).GT.X3(J))X3(J)=X(I,J)
```

```
  1 CONTINUE
    END
```

六、变量标准化的计算结果

样品数＝ 5

变量数＝ 4

变量的标准化方式＝ 1

原始数据表

样品序号	变量1	变量2	变量3	变量4
1	1000.000	500.000	1.500	2000.000
2	250.000	150.000	1.000	2200.000
3	100.000	70.000	3.000	1500.000
4	10.000	200.000	2.000	1800.000
5	40.000	100.000	5.000	2500.000

变换后的数据表

样品序号	变量1	变量2	变量3	变量4
1	.714	.490	.120	.200
2	.179	.147	.080	.220
3	.071	.069	.240	.150
4	.007	.196	.160	.180
5	.029	.098	.400	.250

程序二　变量筛选法

一、程序主要功能

这里介绍一种简单易行的变量筛选方法，可以选出那些有效的变量，剔出那些无效的变量。

二、程序符号说明

N－－－－－－－样品数（要求N小于500）；

M－－－－－－－自变量数（要求M小于49）M＋1为变量总数；

X(500,50)－－原始数据矩阵，矩阵的行号为样品编号，矩阵的列号为变量编号（其中前M列为自变量，第M＋1列为因变量）；

Y(50)－－－－各个变量的门坎值，大于门坎值时令变量为1，小于门坎值时令变量为0；

NX(500,50)－各个变量按门坎值0－－1化后的数据矩阵；

NY(50,50)－－变量的报错相关频数，

其中：NY(M1，50)为合计的报错相关频数；

X1(50)————变量的报对频率；

X2(50)————变量的报错相关频率；

X3(50)————变量的筛选指标，即：X3＝X1/X2；

KX(50) ———筛选后被剔出的变量，令其KX＝1。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为Z3.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
————————————————
N,M
((X(I,J),J=1,M+1),I=1,N)
(Y(J),J=1,M+1)
————————————————
```

例如，下面的数据文件（Z3.DAT）就是一个供用户检测本程序的数据文件：

```
——————————————————————————
10,7
7.8,4.5,1.0,45.0,1.0,0.23,0.0,1.0,
11.9,35.9,1.0,123.9,0.0,0.6,0.0,1.0,
6.3,6.7,1.0,23.0,0.0,0.3,1.0,0.0,
34.7,14.9,0.0,56.7,0.0,0.95,0.0,0.0,
19.4,4.8,1.0,12.6,1.0,0.63,0.0,1.0,
2.0,23.8,1.0,190.6,1.0,0.83,0.0,1.0,
9.9,11.9,0.0,45.6,0.0,0.57,1.0,0.0,
95.7,5.7,1.0,286.4,1.0,0.28,1.0,1.0,
120.8,23.8,0.0,45.9,0.0,0.83,1.0,0.0,
34.0,9.5,0.0,77.4,0.0,0.30,1.0,0.0
10.0,10.0,1.0,100.0,1.0,0.33,1.0,1.0
——————————————————————————
```

四、计算结果输出

本程序输出文件的约定名为Z3.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、变量筛选法程序

```
      PROGRAM Z9003
      COMMON X(200,20),Y(20),NX(200,20),NY(21,20),X1(20),X2(20),
     -X3(20),KX(20)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,30(1H )\)')
```

```
      WRITE(*,'(1X,A)')'变量筛选法'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名〔约定名 Z3.DAT〕:'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z3.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名〔约定名 Z3.WRI〕:'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z3.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入原始数据矩阵; '
      READ(1,*,ERR=20)N,M
      M1=M+1
      READ(1,*,ERR=20)((X(I,J),J=1,M1),I=1,N)
      READ(1,*,ERR=20)(Y(J),J=1,M1)
      WRITE(*,*)'开始进行变量的筛选计算; 请等待! '
      WRITE(2,*)'        * * * * * * * * * * * * * * * * * * * * * *'
      WRITE(2,*)'        *                                         *'
      WRITE(2,*)'        *          变量筛选的计算结果             *'
      WRITE(2,*)'        *                                         *'
      WRITE(2,*)'        * * * * * * * * * * * * * * * * * * * * * *'
      WRITE(2,100)N,M,M1
  100 FORMAT(//5X,'样品数=',I3/5X,'自变量数=',I3/5X,'变量总数=',I3)
      WRITE(2,'(//11X,A)')'原始数据表'
      MXY=M1
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      DO 2 I=1,N
    2 WRITE(2,102)I,(X(I,J),J=1,M1)
  102 FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      WRITE(2,103)'变量门坎值',(Y(J),J=1,M1)
  103 FORMAT(/2X,A,1X,10(F10.3,1X)/13X,10(F10.3,1X))
C     对各个变量按门坎值进行 0--1 化变换，形成 0--1 化的矩阵NX.
      DO 3 J=1,M1
      DO 3 I=1,N
      IF(X(I,J).GE.Y(J))THEN
        NX(I,J)=1
      ELSE
        NX(I,J)=0
```

```
      ENDIF
    3 CONTINUE
      WRITE(2,'(//11X,A)')'变换后的 0——1化矩阵'
      DO 4 I=1,N
    4 WRITE(2,104)(NX(I,J),J=1,M1)
  104 FORMAT(5X,10I5)
C     计算变量的报对频率X1.
      DO 6 J=1,M
      NN=0
      DO 5 I=1,N
      IF(NX(I,J).EQ.NX(I,M1))NN=NN+1
    5 CONTINUE
    6 X1(J)=FLOAT(NN)/FLOAT(N)
C     计算变量的报错相关频数矩阵NY,只计算上三角部分.
      DO 8 I=1,M
      DO 8 J=I,M
      IF(I.EQ.J)GO TO 8
      NN=0
      DO 7 K=1,N
      IF(NX(K,I).EQ.NX(K,J).AND.NX(K,I).NE.NX(K,M1).AND.NX(K,J)
     -.NE.NX(K,M1))NN=NN+1
    7 CONTINUE
      NY(I,J)=NN
    8 CONTINUE
C     按对称关系，形成矩阵NY的下三角部分.
      DO 9 I=2,M
      DO 9 J=1,I-1
    9 NY(I,J)=NY(J,I)
C     输出变量间的报错相关频数矩阵NY.
      WRITE(2,'(//11X,A)')'变量间的报错相关频数矩阵'
      DO 10 I=1,M
   10 WRITE(2,104)(NY(I,J),J=1,M)
C     给剔出变量的标志KX赋零值.
      DO 11 J=1,M
   11 KX(J)=0
C     进行变量筛选计算；MM为筛选后的剩余变量总数.
      MM=M
      DO 18 K=1,M-1
C     计算合计的报错相关频数.
```

```
      DO 12 J=1,M
      NY(M1,J)=0
      DO 12 I=1,M
      IF(KX(I).EQ.1)GO TO 12
      NY(M1,J)=NY(M1,J)+NY(I,J)
   12 CONTINUE
C     计算变量的报错相关频率X2.
      DO 13 J=1,M
      IF(KX(J).EQ.1)GO TO 13
      X2(J)=FLOAT(NY(M1,J))/FLOAT(N*(MM-1))
   13 CONTINUE
C     计算变量的筛选指标X3.
      DO 14 J=1,M
      IF(KX(J).EQ.1)GO TO 14
      IF(X2(J).LT.10E-10)THEN
        X3(J)=10E10
      ELSE
        X3(J)=X1(J)/X2(J)
      ENDIF
   14 CONTINUE
C     输出变量的筛选结果.
      WRITE(2,'(//5X,A,I2,A)')'第',K,'次筛选计算结果如下：'
      WRITE(2,'(/5X,A,6X,A)')'变量号','变量的筛选指标'
      DO 15 I=1,M
      IF(KX(I).EQ.0)THEN
        WRITE(2,105)I,X3(I)
      ELSE
        WRITE(2,106)I
      ENDIF
   15 CONTINUE
  105 FORMAT(9X,I2,F20.3)
  106 FORMAT(9X,I2,12X,'已被剔出')
C     选出起作用最小的变量.
      XMIN=10E15
      DO 16 J=1,M
      IF(KX(J).EQ.1)GO TO16
      IF(X3(J).LT.XMIN)XMIN=X3(J)
   16 CONTINUE
C     挑出要剔掉的变量，令其KX=1，并且剩余变量总数MM-1.
```

```
      DO 17 J=1,M
      IF(KX(J).EQ.1)GO TO 17
      IF(X3(J)-XMIN.LT.10E-10)THEN
        KX(J)=1
        MM=MM-1
      ENDIF
   17 CONTINUE
      IF(MM.EQ.0)GO TO 19
   18 CONTINUE
   19 CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？〔Y/N〕：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
   20 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

六、变量筛选的计算结果

样品数=10

自变量数=7

变量总数=8

原始数据表

样品序号	变量1	变量2	变量3	变量4	变量5	变量6	变量7	变量8
1	7.800	4.500	1.000	45.000	1.000	.230	.000	1.000
2	11.900	35.900	1.000	123.900	.000	.600	.000	1.000
3	6.300	6.700	1.000	23.000	.000	.300	1.000	.000
4	34.700	14.900	.000	56.700	.000	.950	.000	.000
5	19.400	4.800	1.000	12.600	1.000	.630	.000	1.000
6	2.000	23.800	1.000	190.600	1.000	.830	.000	1.000
7	9.900	11.900	.000	45.600	.000	.570	1.000	.000
8	95.700	5.700	1.000	286.400	1.000	.280	1.000	1.000
9	120.800	23.800	.000	45.900	.000	.830	1.000	.000
10	34.000	9.500	.000	77 400	.000	.300	1.000	.000
变量门坎值	10.000	10.000	1.000	100.000	1.000	.330	1.000	1.000

变换后的0—1化矩阵

```
0  0  1  0  1  0  0  1
1  1  1  1  0  1  0  1
```

0 0 1 0 0 0 1 0

1 1 0 0 0 1 0 0

1 0 1 0 1 1 0 1

0 1 1 1 1 1 0 1

0 1 0 0 0 1 1 0

1 0 1 1 1 0 1 1

1 1 0 0 0 1 1 0

1 0 0 0 0 0 1 0

变量间的报错相关频数矩阵

0 3 0 1 0 3 4

3 0 0 2 0 5 4

0 0 0 0 0 0 1

1 2 0 0 0 1 2

0 0 0 0 0 0 1

3 5 0 1 0 0 3

4 4 1 2 1 3 0

第1次筛选计算结果如下：

变量号	变量的筛选指标
1	2.727
2	1.714
3	54.000
4	8.000
5	54.000
6	2.500
7	.800

第2次筛选计算结果如下：

变量号	变量的筛选指标
1	3.571
2	2.000
3	100000000000.000
4	10.000
5	100000000000.000
6	2.778
7	已被剔出

第3次筛选计算结果如下：

变量号	变量的筛选指标
1	5.000
2	已被剔出
3	100000000000.000

4	16.000
5	100000000000.000
6	5.000
7	已被剔出

第4次筛选计算结果如下：

变量号	变量的筛选指标
1	已被剔出
2	已被剔出
3	100000000000.000
4	100000000000.000
5	100000000000.000
6	已被剔出
7	已被剔出

程序三　离群数据的处理

一、程序主要功能

在地质数据中有时出现为数极少的特高（低）值，可比数据的平均值高（低）出很多倍，这种数据称作离群数据或奇异数据。本程序适用于对大子样数据进行离群数据的处理。

二、程序符号说明

N—————— 数据个数（要求N小于5000）；

X(5000)—— 开始为原始数据，之后为正常数据存放单元；

XX(5000)— 异常数据的存放单元；

NN———— 需要处理的数据批数；

N—————— 每批数据的个数；

XL———— 否定域的标准差倍数；

XR———— 否定域的修正系数，即原否定域的增加倍数；

KM———— 直方图的分组数，KM＝5——21，在数据文件中如果令KM＝0时，则KM的值由程序自动生成。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为Z4.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
——————————————
    NN
    N,XL,XR,KM
    (X(I),I=1,N)
——————————————
```

例如，下面的数据文件（Z4.DAT）就是一个供用户检测本程序的数据文件：

```
1
95,2.0,0.5,11
335.80,212.32,151.16,194.92,234.53,159.41,107.14,210.54,
240.12,209.45,256.72,481.15,179.05,269.49,162.18,207.81,
117.50,101.67,155.44,209.68,296.82,213.32,233.24,175.64,
150.01, 92.86, 87.40, 97.00,239.12,254.83,159.44,104.02,
193.52,111.96,104.71,208.03,202.92,143.22,202.14,457.15,
311.24,343.49,217.95,216.67,442.10,283.73,228.37,206.42,
247.15,181.85,216.96,158.14,270.76, 57.97,193.76,  6.38,
297.55,173.29,228.63,240.98,385.99,311.99,222.07,251.99,
179.71,211.46,138.85,179.30,100.98,258.76,183.84, 16.37,
282.06, 99.08,112.11,169.71,187.90, 97.29, 86.63,216.17,
241.88,701.35,380.45,184.07,228.61,169.80,193.76,217.81,
 85.48,296.04,250.89,206.08,194.41,497.58,114.59
```

四、计算结果输出

本程序输出文件的约定名为Z4.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、离群数据处理方法的主程序

```
      PROGRAM Z9004
      DIMENSION X(5000),XX(5000)
      CHARACTER FILENA*30,NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'大子样离群数据处理方法'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名〔约定名 Z4.DAT〕：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z4.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,(1X,A\)')'请输入您的输出文件名〔约定名Z4.WRI〕：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z4.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入原始数据；'
      READ(1,*,ERR=7)NN
      WRITE(*,*)'开始进行离群数据的处理计算，请等待！
      WRITE(2,*)'          ******************'
      WRITE(2,*)'          *                *'
      WRITE(2,*)'          *  离群数据的处理计算结果  *'
```

```
      WRITE(2,*)'        *                    *'
      WRITE(2,*)'        * * * * * * * * * * * * * * * * *'
      WRITE(2,100)NN
  100 FORMAT(//5X,'需要处理的数据批数=',I3)
      DO 6 K=1,NN
      WRITE(2,101)K
  101 FORMAT(//5X,20('*')/5X,'*',18(''),'*'/5X,'*',4(''),'第',I2,
     -'批数据',4(''),'*'/5X,'*',18(''),'*'/5X,20(,*'))
      READ(1,*,ERR=7)N,XL,XR,KM
      WRITE(2,102)N,XL,XR,KM
  102 FORMAT(//5X,'数据个数=',I3/5X,'否定域的标准差倍数=',F5.3/5X,
     -'否定域的修正系数=',F5.3/5X,'直方图的分组数=',I3)
      READ(1,*,ERR=7)(X(I),I=1,N)
      WRITE(2,103)NN,N
  103 FORMAT(//5X,'第',I3,'批数据(总共',I5,'个)')
      WRITE(2,104)(X(I),I=1,N)
  104 FORMAT(5X,10F12.3)
      CALL X1S(N,X,X1,S)
      CALL MAP(N,X,KM)
C     下面的NS为样品总数，M为处理的次数，KS2为各次被剔出的离群数据累计数，
      NS=N
      M=0
      K2S=0
C     下面的XP1为离群数据的上限值，XP2为离群数据的下限值.
 1000 XP1=X1+S*XL
      XP2=X1-S*XL
C     下面的U1为离差的三次方和，   U2为离差的四次方和；
C                R1为三阶中心矩，     R2为四阶中心矩；
C            P1为三阶中心矩否定域，P2为四阶中心矩否定域.
      U1=0.0
      U2=0.0
      DO 2 I=1, N
      XS=X(I)-X1
      U1=U1+XS*XS*XS
    2 U2=U2+U1*XS
      U1=U1/N
      U2=U2/N
      R1=U1/S**3
      R2=(U2/S**4)-3
```

```
      P1=1.96*SQRT(6.0/N)
      P2=1.96*SQRT(24.0/N)
      P1=P1+P1*XR
      P2=P2+P2*XR
C     当三阶中心矩大于否定域，或者四阶中心矩大于否定域时，继续进行剔出离群数据
C     的计算；否则计算结束.
      IF(ABS(R1).GT.P1.OR.ABS(R2).GT.P2)GO TO 3
      GO TO 6
C     当处理的次数大于5次时，则计算结束.
    3 IF(M.GE.5)GO TO 6
C     下面的K1为保留的正常数据个数，K2为被剔出的离群数据个数；
      K1=0
      K2=0
      DO 5 I=1,N
      IF(X(I).GT.XP1.OR.X(I).LT.XP2)GO TO 4
      K1=K1+1
      X(K1)=X(I)
      GO TO 5
    4 K2=K2+1
      XX(K2)=X(I)
    5 CONTINUE
C     各次被剔出的离群数据累计相加.
      K2S=K2S+K2
C     把保留的正常数据个数K1作为再次处理时的样品总数.
      N=K1
C     处理的次数增加1.
      M=M+1
C     输出处理的次数M.
      WRITE(2,105)M
  105 FORMAT(//5X,20('—')/5X,'I',18('  '),'I'/5X,'I',4('  '),'第',
     I2,—'次处理',4('  '),'I'/5X,'I',18('  '),'I'/5X,20('—'))
C     输出数据的平均值X1及标准差S.
      WRITE(2,106)X1,S
  106 FORMAT(//5X,'平均值=',F12.3/5X,'标准差=',F12.3)
C     输出离群数据的上，下限值XP1,XP2.
      WRITE(2,107)XP1,XP2
  107 FORMAT(//5X,'离群数据的上限值=',F12.3/5X,'离群数据的下限值=',
     —F12.3)
C     输出保留的正常数据个数K1及数据.
```

```
      IF(K1.NE.0)WRITE(2,108)K1,(X(I),I=1,K1)
  108 FORMAT(//5X,'保留的正常数据(共有',I3,'个): '/(5X'10F12.3))
C     输出剔出的离群数据个数K2及数据.
      IF(K2.NE.0)WRITE(2,109)K2,(XX(I),I=1,K2)
  109 FORMAT(//5X,'剔出的离群数据(共有',I3,'个): '/(5X,10F12.3))
C     计算并输出下面的 PK1为保留的正常数据个数占数据总数的百分数;
C                      PK2为剔出的离群数据个数占数据总数的百分数;
C                 PK2S为剔出的离群数据累计个数占数据总数的百分数.
      PK1=FLOAT(K1)/FLOAT(NS)*100.0
      PK2=FLOAT(K2)/FLOAT(NS)*100.0
      PK2S=FLOAT(K2S)/FLOAT(NS)*100.0
      WRITE(2,110)K1,PK1,K2,PK2,K2S,PK1S
  110 FORMAT(//
     -5X,'保留的正常数据个数=',I3,5X.,占数据总数的百分数=',F10.3/
     -5X,'剔出的离群数据个数=',I3,5X,'占数据总数的百分数=',F10.3/
     -1X,'剔出的离群数据累计个数=',I3,5X,'占数据总数的百分数=',F10.3)
C     调用计算标准差子程序.
      CALL X1S(N,X,X1,S)
C     调用宽行打图子程序,打印原始数据的直方图.
      CALL MAP(N,X,KM)
C     转向标号1000,再次进行离群数据处理.
      GO TO 1000
    6 CONTINUF
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭!还继续进行计算吗? (Y/N):'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    7 WRITE(*,*)'您的数据文件有错!'
      STOP
      END
```

(1)计算标准差子程序

```
      SUBROUTINE X1S(N,X'X1,S)
      DIMENSIONX(N)
      SS=0.0
      DO 1 I=1,N
    1 SS=SS+X(I)
      X1=SS/N
```

```
      SS=0.0
      DO 2 I=1,N
      DX=X(I)-X1
    2 SS=SS+DX*DX
      S=SQRT(1.0/(N-1)*SS)
      END
```

(2)宽行打图子程序

```
      SUBROUTINE MAP(N,X,KM)
      DIMENSION X(N),XH(22),NX(21),P(21),PP(105),H(26),W(106)
      CHARACTER W*1,FM*100
      DATA A,B,C,D/' ','*','I','+'/
      IF(KM.EQ.0)THEN
        M=N/5
        IF(MOD(M,2).EQ.0)M=M-1
      ELSE
        M=KM
      ENDIF
      IF(M.GT.21)M=21
      IF(M.LT.5)M=5
      M1=M+1
      XMIN=X(1)
      XMAX=X(1)
      DO 1 I=2,N
      IF(X(I).LT.XMIN)XMIN=X(I)
      IF(X(I).GT.XMAX)XMAX=X(I)
    1 CONTINUE
      XL=XMAX-XMIN
      DX=XL/FLOAT(M)
      WRITE(2,101)XMIN,XMAX,XL,DX
  101 FORMAT(//5X,'数据中的极小值=',F12.3/5X,'数据中的极大值=',F
     -12.3/5X,'数据的极差=',F12.3/5X,'分组区间值=',F12.3)
      DO 2 I=1,M1
    2 XH(I)=XMIN+DX*(I-1)
      XH(1)=XH(1)-DX
      DO 5 K=1,M1
      NX(K)=0
      DO 4 I=1,N
      IF(X(I).GT.XH(K).AND.X(I).LE.XH(K+1))GO TO 3
      GO TO 4
```

```
3 NX(K)=NX(K)+1
4 CONTINUE
5 P(K)=FLOAT(NX(K))/FLOAT(N)
  XH(1)=XH(1)+DX
  K=1
  MM=M*5
  MM1=MM+1
  DO 6 I=1,MM
  PP(I)=P(K)
  IF((I-I/5*5).EQ.0)K=K+1
6 CONTINUE
  DO 7 I=1,MM1,5
7 W(I)=A
  DO 8 I=1,26
8 H(I)=100.0-FLOAT((I-1)*4)
  WRITE(2,'(//2,X,A/11X,A)')'数据的直方图','概率'
  WRITE(FM,'(A10,A4,I3,A9,A4,I3,A6)')'(5X,F10.2,','1H+,',M,
 -'(5H-----+)','/8X,',M1,'F10.2)'
  DO 13 I=1,25
  W(1)=C
  DO 9 J=2,MM1
  IF((J-J/5*5).EQ.1)GO TO 9
  W(J)=A
9 CONTINUE
  DO 11 J=1,MM
  IF(100.0*PP(J).LE.H(I).AND.100.0*PP(J).GT.H(I+1))GO TO 10
  GO TO 11
10 W(J)=B
  W(J+1)=B
11 CONTINUE
  IF(IFIX(H(I))/10*10-IFIX(H(I)).EQ.0)GO TO 12
  WRITE(2,'(15X,110A1)')(W(J),J=1,MM1)
  GO TO 13
12 W(1)=D
  WRITE(2,'(5X,F10.2,110A1)')H(I),(W(J),J=1,MM1)
13 CONTINUE
  WRITE(2,FM)H(26),(XH(I),I=1,K,2)
  WRITE(2,'(13X,11F10.2)')(XH(I),I=2,K,2)
  END
```

六、离群数据处理的计算结果

需要处理的数据批数＝1

第1批数据：

数据个数＝95

否定域的标准差倍数＝2.000

否定域的修正系数＝.500

直方图的分组数＝11

第1批数据(总共　95个)

335.800	212.320	151.160	194.920	234.530
159.410	107.140	210.540	240.120	209.450
256.720	481.150	179.050	269.490	162.180
207.810	117.500	101.670	155.440	209.680
296.820	213.320	233.240	175.640	150.010
92.860	87.400	97.000	239.120	254.830
159.440	104.020	193.520	111.950	104.710
208.030	202.920	143.220	202.140	457.150
311.240	343.490	217.950	216.670	442.100
283.730	228.370	206.420	247.150	181.850
216.960	158.140	270.760	57.970	193.760
6.380	297.550	173.290	228.630	243.980
385.990	311.990	222.070	251.990	179.710
211.460	138.850	179.300	100.980	258.760
183.840	16.370	282.060	99.080	112.110
169.710	187.900	97.290	86.530	216.170
241.880	701.350	380.450	184.070	228.610
169.800	193.760	217.810	85.480	296.040
250.890	206.080	194.410	497.580	114.590

数据中的极小值＝　6.380

数据中的极大值＝　701.350

数据的极差＝　694.970

分组区间值＝　63.179

第1次处理：

平均值＝　211.629

标准差＝　104.171

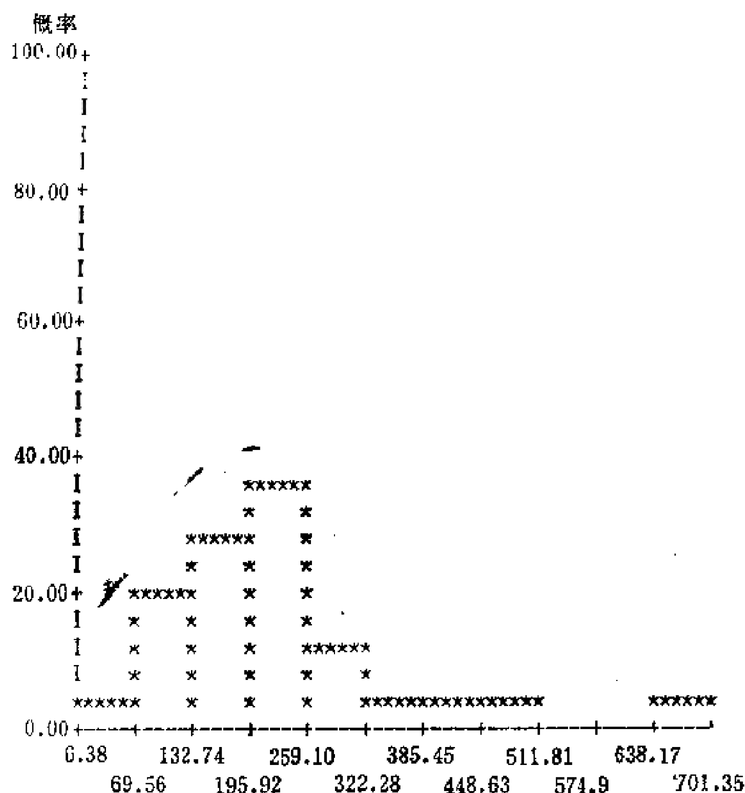
离群数据的上限值＝　419.972

离群数据的下限值＝　3.287

保留的正常数据(共有90个)：

335.800	212.320	151.160	194.920	234.520
159.410	107.140	210.540	240.120	209.450

256.720	179.050	269.490	162.180	207.810
117.500	101.670	155.440	209.680	296.820
213.320	233.240	175.640	150.010	92.860



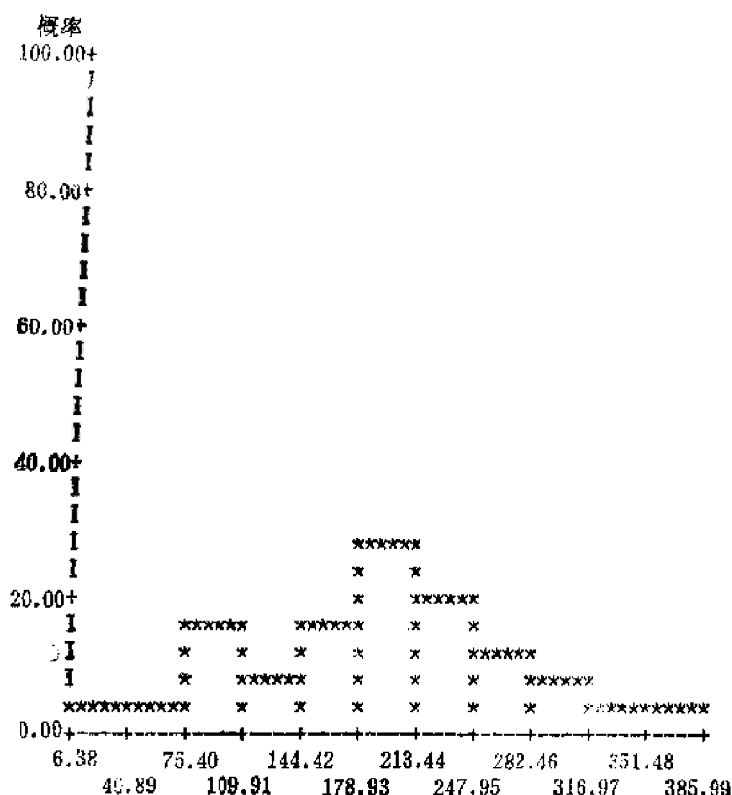
附图1 原始数据的直方图

87.400	97.000	239.120	254.830	159.440
104.020	193.520	111.950	104.710	208.030
202.920	143.220	202.140	311.240	343.490
217.950	216.670	283.730	228.370	206.420
247.150	181.850	216.960	158.140	270.760
57.970	193.760	6.380	297.550	173.290
228.630	243.980	385.990	311.990	222.070
251.990	179.710	211.460	138.850	179.300
100.980	258.760	183.840	16.370	282.060
99.080	112.110	169.710	187.900	97.290
86.530	216.170	241.880	380.450	184.070
228.610	169.800	193.760	217.810	85.480
296.040	250.890	206.080	194.410	114.590

剔出的离群数据（共有 5个）：

481.150	457.150	442.100	701.350	497.580

保留的正常数据个数＝ 90　　占数据总数的百分数＝ 94.737
剔出的离群数据个数＝ 5　　占数据总数的百分数＝ 5.263
剔出的离群数据累计个数＝ 5　　占数据总数的百分数＝ 5.263
数据中的极小值＝ 6.380
数据中的极大值＝ 385.990
数据的极差＝ 379.610
分组区间值＝ 34.510



附图2　处理后的直方图

程序四　绘制随机变量直方图

一、程序主要功能

本程序可对一组随机变量进行分组整理，求其平均值，方差，标准差，极差，偏倚系数等特征参数，并且可以绘制密度分布直方图、分布函数直方图。

二、程序符号说明

NN —————需要处理的数据批数；

N ——————每批的数据个数（要求N小于5000）；

X(5000)——原始数据；

XH(22) ——直方图中的分组间隔值（横坐标）；

NX(21) ——各组的频数；

P(21)———各组的频率；

F(21)———各组的累计频率；

H(26)———直方图中的概率间隔值（纵坐标）；

W(106)——————打图时的字符存放单元；

KM ———————直方图的分组数，KM=5−21，在数据文件中如果令KM=0时，则KM的值由程序自动生成，如果分组数为偶数时，则自动减1，使分组数为奇数，以便于显示峰值。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为Z5.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
NN
N,KM
(X(I,J),I=1,N)
```

例如，下面的数据文件（Z5.DAT）就是一个供用户检测本程序的数据文件：

```
1
95,0
285.07,181.68,127.80,167.17,205.69,134.95, 90.77,180.06,
204.26,183.28,220.74,419.53,154.75,239.71,136.03,180.18,
 98.42, 84.92,129.72,171.04,236.54,165.07,188.95,140.96,
118.19, 71.24, 69.92, 80.25,193.76,219.70,115.71, 81.44
164.66, 91,96, 83.86,183.17,157.32,120.21,164.74,368.64,
253.10,275.54,175.34,168.55,347.60,232.70,187.31,172.65,
204.44,149.99,172.29,130.77,208.87, 41.62,153.42,  4.65,
221.50,135.75,183.59,217.63,322.65,281.05,197.15,223.93,
154.92,178.52,118.73,151.57, 89.49,226.08,160.94, 14.41,
245.30, 83.43, 95.67,147.77,163.32, 81.30, 72.99,177.58,
201.39,522.61,314.21,151.43,203.24,152.23,171.64,195.67,
 77.90,266.53,222.83,183.95,172.24,442.42,101.02
```

四、计算结果输出

本程序输出文件的约定名为Z5.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、绘制随机变量直方图的主程序

```
      PROGRAM Z9005
      COMMON X(5000),XH(22),NX(21),P(21),F(21),PP(105),FF(105),
```

```
      H(26),W(106)
      CHARACTER FILENA*30,NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X, A)')'绘制随机变量的直方图'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名Z5.DAT]: '
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z5.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名[约定名Z5.WRI]: '
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z5.WRI'
      OPEN(2,FILE=FILENA, STATUS='NEW')
      WRITE(*,*)'开始读入原始数据; '
      READ(1,*,ERR=3)NN
      WRITE(*,*)'开始绘制随机变量直方图; 请等待! '
      WRITE(2, *)'         * * * * * * * * * * * * * * * * * * *'
      WRITE(2, *)'         *                                   *'
      WRITE(2, *)'         *  绘制随机变量直方图的计算结果       *'
      WRITE(2, *)'         *                                   *'
      WRITE(2, *)'         * * * * * * * * * * * * * * * * * * *'
      WRITE(2,100)NN
  100 FORMAT(//5X,'需要处理的数据批数=', I3 )
      DO 2 K=1,NN
      WRITE(2, 101)K
  101 FORMAT(//5X,20('*')/5X,'*',18('  '),'*'/5X,'*',4('  '),'第',
     -I2,'批数据',4('  '),'*'/5X,'*',18('  '),'*'/5X,20('*'))
      READ(1,*,ERR=3)N, KM
      WRITE(2,102)N,KM
  102 FORMAT(//5X,'数据个数=',I3/5X,'直方图的分组数=',I3)
      READ(1,*,ERR=3)(X(I),I=1,N)
      WRITE(2,103)NN,N
  103 FORMAT(//5X,'第', I3, '批数据(总共', I5, '个)')
      WRITE(2,104)(X(I),I=1,N)
  104 FORMAT(5X,10F12.3)
      CALL X1S(N)
      CALL MAP(N,KM)
    2 CONTINUE
      CLOSE(1)
      CLOSE(2)
```

```
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    3 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

（1）计算随机变量特征值的子程序

```
      SUBROUTINE X1S(N)
      COMMON X(5000),XH(22),NX(21),P(21),F(21),PP(105),FF(105),
     -H(26),W(106)
      SS=0.0
      DO 1 I=1,N
    1 SS=SS+X(I)
      X1=SS/N
      SS=0.0
      DO 2 I=1,N
      DX=X(I)-X1
    2 SS=SS+DX*DX
      S2=(1.0/(N-1))*SS
      S=SQRT(S2)
      C=S/X1
      WRITE(2,101)X1,S2,S,C
  101 FORMAT(//5X,'平均值=',F12.3/5X,'方差=',F12.3/5X,'标准差=',
     -F12.3/5X,'偏倚系数=',F12.3)
      END
```

（2）宽行打印字符图子程序

```
      SUBROUTINE MAP(N, KM)
      COMMON X(5000), XH(22), NX(21), P(21), F(21), PP(105),
     -FF(105), H(26), W(106)
      CHARACTER W*1, FM*100
      DATA A, B, C,D/' ','*','I','+'/
      IF(KM.EQ.0)THEN
        M=N/5
        IF(MOD(M, 2).EQ.0)M=M-1
      ELSE
        M=KM
      ENDIF
      IF(M.GT.21)M=21
```

```
      IF(M.LT.5)M=5
      M1=M+1
      XMIN=X(1)
      XMAX=X(1)
      DO 1 I=2,N
      IF(X(I).LT.XMIN)XMIN=X(I)
      IF(X(I).GT.XMAX)XMAX=X(I)
    1 CONTINUE
      XL=XMAX-XMIN
      DX=XL/FLOAT(M)
      WRITE(2,101)XMIN, XMAX, XL, DX
  101 FORMAT(//5X,'数据中的极小值=', F12.3/5X, '数据中的极大值=',
     -F12.3/5X, '数据的极差=', F12.3/5X, '分组区间值=', F12.3)
      DO 2 I=1,M1
    2 XH(I)=XMIN+DX*(I-1)
      WRITE(2, 102)M1,(XH(I), I=1,M1)
  102 FORMAT(//10X,'分组区间值(总计', I2, '个值)'/(5X, 10F12.3))
      XH(1)=XH(1)-DX
      S=0.0
      DO 5 K=1,M1
      NX(K)=0
      DO 4 I=1, N
      IF(X(I).GT.XH(K).AND.X(I).LE.XH(K+1))GO TO 3
      GO TO 4
    3 NX(K)=NX(K)+1
    4 CONTINUE
      P(K)=FLOAT(NX(K))/FLOAT(N)
      F(K)=1.0-S
    5 S=S+P(K)
      XH(1)=XH(1)+DX
      WRITE(2, 103)M, (P(I), I=1,M)
  103 FORMAT(//5X,'各组的频率值(总计', I2, '个值)'/(5X, 10F12.3))
      WRITE(2,104)M,(F(I),I=1,M)
  104 FORMAT(//5X,'各组的累计频率值(总计',I2,'个值)'/(5X, 10F12.3))
      K=1
      MM=M*5
      MM1=MM+1
      DO 6 I=1,MM
      PP(I)=P(K)
```

```
  FF(I)=F(K)
  IF((I-I/5*5).EQ.0)K=K+1
6 CONTINUE
  DO 7 I=1, MM1, 5
7 W(I)=A
  DO 8 I=1, 26
  8 H(I)=100.0-FLOAT((I-1)*4)
  WRITE(2,'(//20X, A/11X,A)')'密度分布直方图', '概率'
  WRITE(FM,'(A10, A4, I3, A9, A4, I3, A6)')'(5X,F10.2,','1H+,
-', M, '(5H----+)',','/8X,',M1, 'F10.2)'
  DO 13 I=1, 25
  W(1)=C
  DO 9 J=2, MM1
  IF((J-J/5*5).EQ.1)GO TO 9
  W(J)=A
9 CONTINUE
  DO 11 J=1, MM
  IF(100.0*PP(J).LE.H(I).AND.100.0*PP(J).GT.H(I+1))GO TO 10
  GO TO 11
10 W(J)=B
  W(J+1)=B
11 CONTINUE
  IF(IFIX(H(I))/10*10-IFIX(H(I)).EQ.0)GO TO 12
  WRITE(2,'(15X,110A1)')(W(J),J=1,MM1)
  GO TO 13
12 W(1)=D
  WRITE(2,'(5X, F10.2, 110A1)')H(I), (W(J), J=1, MM1)
13 CONTINUE
  WRITE(2, FM)H(26), (XH(I), I=1, K, 2)
  WRITE(2,'(13X, 11F10.2)')(XH(I), I=2, K, 2)
  WRITE(2,'(//20X,A/11X,A)')'分布函数直方图', '概率'
  DO 14 I=1, MM1, 5
14 W(I)=A
  DO 19 I=1, 25
  W(1)=C
  DO 15 J=2, MM1
  IF((J-J/5*5).EQ.1)GO TO 15
  W(J)=A
15 CONTINUE
```

```
      DO 17 J=1,MM
      IF(100.0*FF(J).LE.H(I).AND.100.0*FF(J).GT.H(I+1))GO TO 16
      GO TO 17
   16 W(J)=B
      W(J+1)=B
   17 CONTINUE
      IF(IFIX(H(I))/10*10-IFIX(H(I)).EQ.0)GO TO 18
      WRITE(2,'(15X,110A1)')(W(J),J=1,MM1)
      GO TO 19
   18 W(1)=D
      WRITE(2,'(5X,F10.2, 110A1)')H(I), (W(J), J=1, MM1)
   19 CONTINUE
      WRITE(2,FM)H(26), (XH(I), I=1, K, 2)
      WRITE(2,'(13X,11F10.2)')(XH(I), I=2, K, 2)
      END
```

六、绘制随机变量直方图的计算结果

需要处理的数据批数=1

第1批数据:

数据个数=95

直方图的分组数=0

第1批数据（总共95个）

285.070	181.680	127.860	167.170	205.690
134.950	90.770	180.060	204.260	183.280
220.740	419.530	154.750	239.710	136.030
180.180	98.420	84.920	129.720	171.040
236.540	165.070	188.950	140.960	118.190
71.240	69.920	80.250	193.760	219.700
115.710	81.440	164.660	91.960	83.860
183.170	157.320	120.210	164.740	368.640
253.100	275.540	175.340	168.550	347.600
232.700	187.310	172.650	204.440	149.990
172.290	130.770	208.870	41.620	153.420
4.650	221.500	135.750	183.590	217.630
322.650	281.050	197.150	223.930	154.920
178.520	118.730	151.570	89.490	226.080
160.940	14.410	245.300	83.430	95.670
147.770	163.320	81.300	72.990	177.580
201.390	522.610	314.210	151.430	203.240
152.230	171.640	195.670	77.900	266.530

222.830　　183.950　　172.240　　442.420　　101.020

平均值＝177.005

方差＝7254.709

标准差＝85.175

偏倚系数＝.481

数据中的极小值＝　4.650

数据中的极大值＝522.610

数据的极差＝517.960

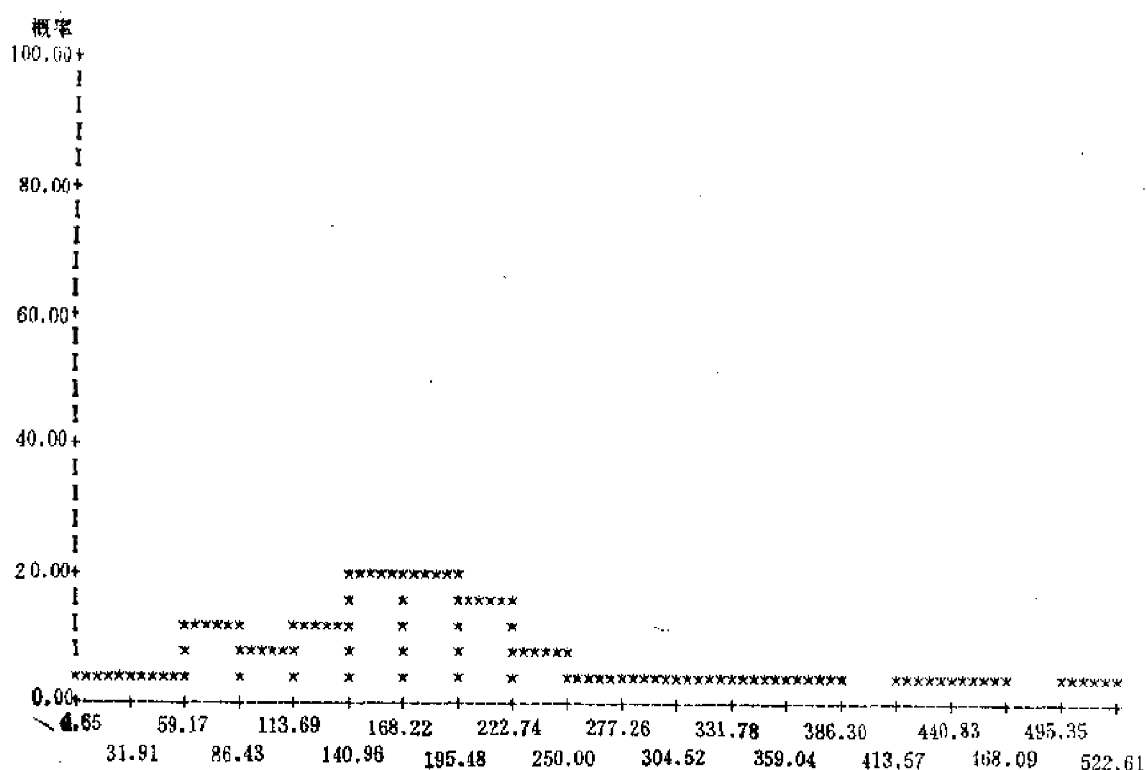
分组区间值＝ 27.261

分组区间值（总计20个值）

4.650	31.911	59.172	86.433	113.694
140.955	168.216	195.477	222.738	249.999
277.260	304.522	331.783	359.044	386.305
413.566	440.827	468.088	495.349	522.610

各组的频率值（总计19个值）

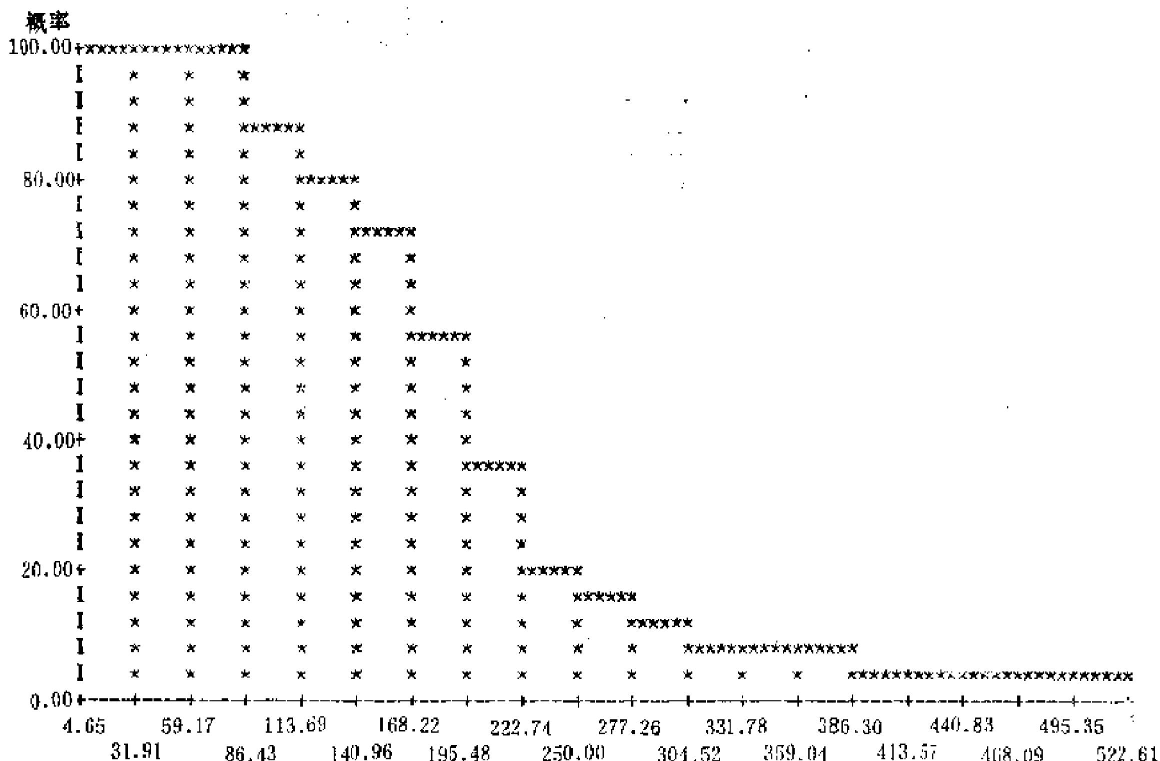
.021	.011	.105	.063	.105
.168	.200	.126	.074	.032
.021	.021	.011	.011	.000
.011	.011	.000	.011	



附图3　密度分布直方图

各组的累计频率值（总计19个值）

1.000	.979	.968	.863	.800
.695	.526	.326	.200	126
.095	.074	.053	.042	.032
032	021	011	011	



附图4　分布函数直方图

程序五　打印三角坐标图

一、程序主要功能

本程序可以将具有三个变量的一组样品打印成三角坐标图。

二、程序符号说明

NN ———————需要处理的数据批数；

XYZ(1000,3) —原始数据矩阵（三个变量应为闭合变量）；

X(122) —————X方向的坐标；

Y(62)——————Y方向的坐标；

N ————————每批数据中的数据个数；

NS ———————每批数据中的数据种类；

KK ———————三角图底边的字符数（KK=40——120）。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为Z6.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
----------------------
NN
NS, KK
N
((XYZ(I, J), J=1, 3), I=1, N)
----------------------
```

例如，下面的数据文件（Z6.DAT）就是一个供用户检测本程序的数据文件：

```
----------------------
1
3, 100
6
0.0, 0.0,100.0,
5.0, 5.0, 90.0,
10.0,10.0,80.0,
15.0,15.0,70.0,
20.0,20.0,60.0,
25.0,25.0,50.0
3
30.0,30.0,40.0,
35.0,35.0,30.0,
40.0,40.0,20.0
2
45.0,45.0,10.0,
50.0, 50.0, 0.0
----------------------
```

四、计算结果输出

本程序输出文件的约定名为Z6.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、打印三角坐标图的主程序

```
      PROGRAM Z9006
      COMMON XYZ(1000,3), X(122), Y(62), W(61,121)
      CHARACTER FILENA*30, W, NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'绘制三角坐标图的方法'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名Z6.DAT]: '
```

```
      READ( *,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='Z6.DAT'
      OPEN(1, FILE=FILENA)
      WRITE( *,'(1X,A\)')'请输入您的输出文件名[约定名Z6.WRI]: '
      READ( *,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='Z6.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE( *, *)'开始读入绘制三角图的原始数据矩阵; '
      READ(1, *, ERR=5)NN
      WRITE( *, * )'开始绘制三角图; 请等待; '
      WRITE(2, * )'        * * * * * * * * * * * * * * * *'
      WRITE(2, * )'        *                             *'
      WRITE(2, * )'        *    绘制三角图的计算结果     *'
      WRITE(2, * )'        *                             *'
      WRITE(2, * )'        * * * * * * * * * * * * * * * *'
      WRITE(2, 100)NN
100   FORMAT(//5X,'需要处理的数据批数=',I3)
      DO 4 K=1,NN
      WRITE(2,101)K
101   FORMAT(//5X,20('*')/5X,'*',18(' '),'*'/5X,'*',4(' '),'第',
     -I2,'批数据',4(' '),'*'/5X,'*',18(' '),'*'/5X,20('*'))
      READ(1, *,ERR=5)NS,KK
      WRITE(2,102)NS,KK
102   FORMAT(//5X,'本批数据的数据种类=',I3/5X,'三角图底边的字符数=',
     -I3)
      CALL COMP0(KK)
      DO 3 L=1,NS
      READ(1, *,ERR=5)N
      READ(1, *,ERR=5)((XYZ(I,J),J=1,3),I=1,N)
      WRITE(2,'(//11X,A,I2,A)')'第', L, '种原始数据表'
      WRITE(2,'(/5X,A,3(4X,A,1X))')'样品序号','变量X','变量Y','变量Z'
      DO 2 I=1,N
    2 WRITE(2,103)I,(XYZ(I,J),J=1,3)
103   FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
    3 CALL COMP1(N,L,KK)
      CALL COMP2(KK)
    4 CONTINUE
      CLOSE(1)
      CLOSE(2)
```

```
      WRITE( *,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    5 WRITE( *, *)'您的数据文件有错！'
      STOP
      END
```

（1）绘制三角坐标图底图的子程序

```
      SUBROUTINE COMP0(KK)
      COMMON XYZ(1000,3),X(122),Y(62),W(61,121)
      CHARACTER*1 W
      DATA A,B,C/' ','*','.'/
      II=KK/2
      JJ=KK+1
      DO 4 I=1, II
      I1=I-1
      N1=MOD(I1, 5)
      J0=II+1-I
      J1=J0+1
      J2=II+I
      J3=J2-J1
      DO 1 J=1,JJ
    1 W(I,J)=A
      IF(N1.NE.0)THEN
        DO 2 J=1, J3, 10
        J4=J0+J
        J5=J4+N1*2
        W(I, J4)=C
    2   W(I, J5)=C
        W(I, J1)=B
        W(I, J2)=B
      ELSE
        DO 3 J=J1, J2, 2
    3   W(I, J)=C
        W(I, J1)=B
        W(I, J2)=B
      ENDIF
    4 CONTINUE
      III=II/5
```

```
      DO 5 I=1, II, III
      J1=II-I
      J2=II+3+I
      W(I, J1)='0'
    5 W(I, J2)='0'
      W(1, II-3)='1'
      W(1, II-2)='0'
      W(1, II+3)='0'
      W(III+1, II-III-2)='8'
      W(III+1, II+III+3)='2'
      W(III*2+1, II-III*2-2)='6'
      W(III*2+1, II+III*2+3)='4'
      W(III*3+1, II-III*3-2)='4'
      W(III*3+1, II+III*3+3)='6'
      W(III*4+1, II-III*4-2)='2'
      W(III*4+1, II+III*4+3)='8'
      DO 6 J=1, JJ
    6 W(II+1, J)=A
      DO 7 J=1, JJ, 2
    7 W(II+1, J)=B
      END
```

（2）确定数据在图中位置的子程序

```
      SUBROUTINE COMP1(N, L, KK)
      COMMON XYZ(1000, 3), X(122), Y(62), W(61, 121)
      CHARACTER*1 W, R(10)
      DATA R/'A','B','C','D','E','F','G','H','I','J'/
      II=KK/2+1
      II1=II+1
      JJ=KK+1
      JJ1=JJ+1
      KK1=KK/2
      DO 1 I=1,II1
    1 Y(I)=100.0-(100.0/KK1)*(I-1)
      DO 2 J=1, JJ1
    2 X(J)=(100.0/KK)*(J-1)
      Y(1)=Y(1)+1
      X(1)=X(1)-1
      DO 6 I=1,II
      J1=II-I
```

```
      DO 6 K=1,N
      IF(XYZ(K,1).LE.Y(I).AND.XYZ(K,1).GT.Y(I+1))GO TO 3
      GO TO 6
    3 DO 5 J=1,JJ
      IF(XYZ(K,2).GE.X(J).AND.XYZ(K,2).LT.X(J+1))GO TO 4
      GO TO 5
    4 W(I, J+J1)=R(L)
      GO TO 6
    5 CONTINUE
    6 CONTINUE
      Y(1)=Y(1)-1
      X(1)=X(1)+1
      END
```

（3）绘制三角坐标图边框的子程序

```
      SUBROUTINE COMP2(KK)
      COMMON XYZ(1000, 3), X(122), Y(62), W(61, 121)
      CHARACTER *1 W, FM *100
      K=KK/10-3
      II=KK/2
      II1=II+1
      II6=II+6
      JJ=KK+1
      LL=K *2+4
      LL1=LL+1
      WRITE(FM, '(A3,I3,A1,A15)' )'(//',II1,'X',',10H三角坐标图)'
      WRITE(2, FM)
      WRITE(FM,'(A2,I3,A1,A5)' )'(/',II6,'X',',1HY)'
      WRITE(2, FM)
      DO 1 I=1, II
    1 WRITE(2,'(6X,121A1)')(W(I,J),J=1, JJ)
      WRITE(FM,'(A11,I3,A2,A10)' )'(2X,4HZ 0,',JJ,'A1',',6H 100X)'
      WRITE(2, FM)(W(II+1, J),J=1, JJ)
      WRITE(FM,'(A10,I2,A5,A6,I2,A5,A6,I2,A5,A6,I2,A5,A6,I2,A5,
     - A5)' )'(4X,3H100,',LL,'(1H  )',',2H80,',LL,'(1H  )',',2H60,',
     - LL,'(1H  )',',2H40,',LL',(1H  )',',2H20,',LL1,'(1H  )',',1H0)'
      WRITE(2, FM)
      END
```

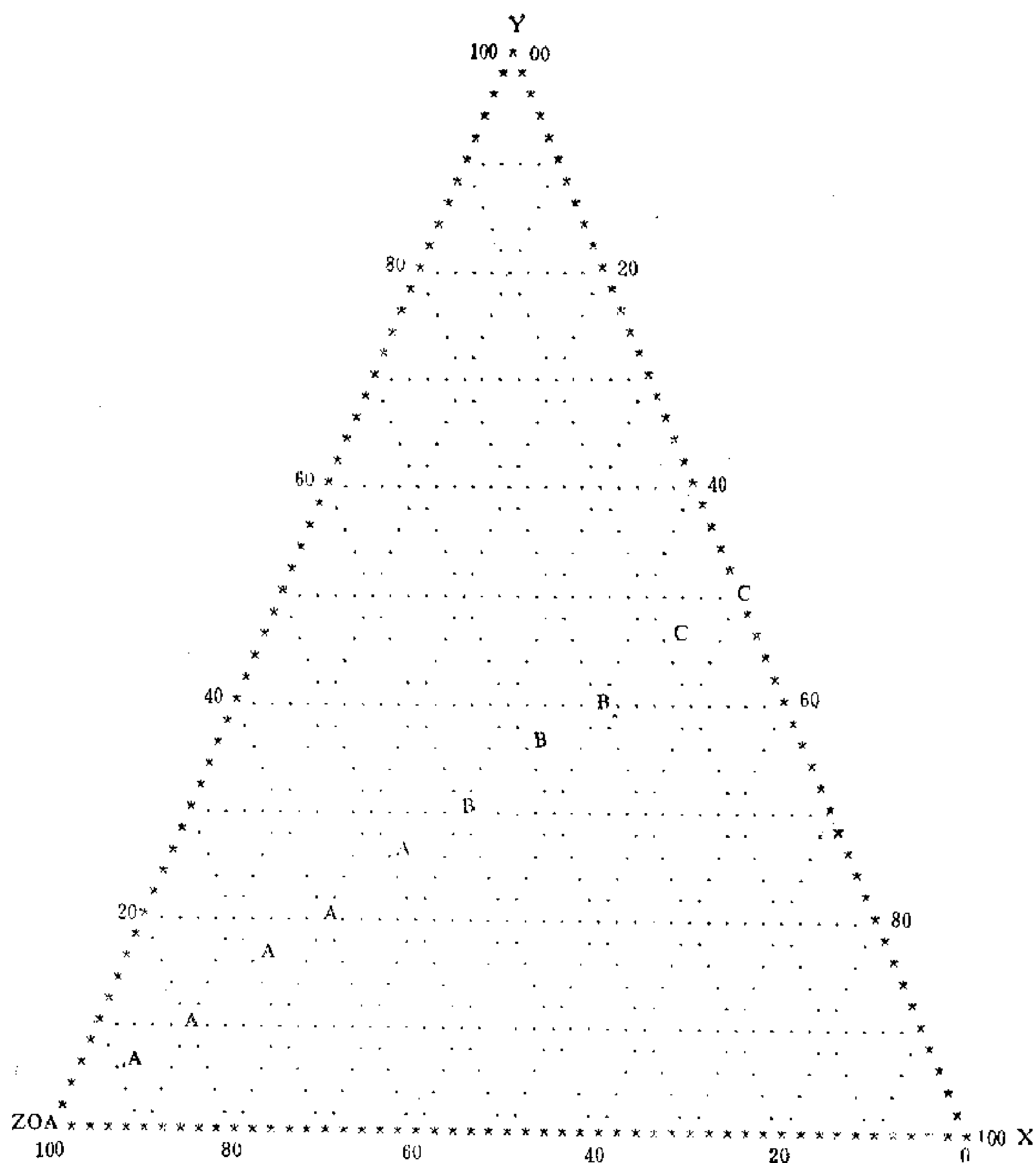
六、打印三角图的计算结果

需要处理的数据批数=1

第1批数据：

本批数据的数据种类＝3

三角图底边的字符数＝100



附图5 三角坐标图

第1种原始数据表

样品序号	变量X	变量Y	变量Z
1	.000	.000	100.000
2	5.000	5.000	90.000
3	10.000	10.000	80.000
4	15.000	15.000	70.000
5	20.000	20.000	60.000
6	25.000	25.000	50.000

第2种原始数据表

样品序号	变量X	变量Y	变量Z
1	30.000	30.000	40.000
2	35.000	35.000	30.000
3	40.000	40.000	20.000

第3种原始数据表

样品序号	变量X	变量Y	变量Z
1	45.000	45.000	10.000
2	50.000	50.000	.000

程序六　一元线性回归分析

一、程序主要功能

一元线性回归分析是研究因变量Y与一个自变量X之间的线性相关关系的一种统计分析方法。

（1）本程序可以根据已知的n对观测数据（Xi，Yi）（i＝1，2，…，N）建立因变量Y与自变量X之间的一元线性回归方程，即：Y＝A＋BX．方程中的待定系数是按最小二乘法计算的；

（2）本程序的计算结果存入输出文件中，可由宽行打印机输出，其中包括：原始数据表，一元线性回归方程式及图形等，其中的预测上、下限方程是按两倍标准差估计的预测值范围，即：对于每个X值，其对应的Y值出现在上、下限方程之间的可能性为95％；

（3）所建立的回归方程是否有意义，可用相关系数R进行检验，如果R大于置信水平下的检验值，则认为回归方程有意义，否则，认为方程无意义；

（4）对于新的未知样品，可将X值代入回归方程Y＝A＋BX求出Y的估计值。

二、程序符号说明

N－－－－－－－样品数；

XY(500,2)－－原始数据矩阵，矩阵的行号为样品编号，列号为变量编号，自变量放在第一列，因变量放在第二列；

X(51)－－－－宽行输出图形的X轴刻度；

Y(26)－－－－宽行输出图形的Y轴刻度；

W(50)－－－－图形输出的字符单元；

YY0(26)－－－回归方程的内存单元；

YY1(26)———预测上限方程的内存单元；

YY2(26)———预测下限方程的内存单元；

FILENA*20—文件字符；

A——————— 回归方程的截距；

B——————— 回归方程的斜率；

R——————— 相关系数。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D1—1.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
N, NN
((XY(I,J),J=1,2),I=1,N)
[(XY(I,1),I=N+1,N+NN+1)]
```

注：方括号[ ]中的数据为选择项

例如，下面的数据文件（D1—1.DAT）就是一个供用户检测本程序的数据文件：

```
11, 2
40.25, 40.16,
46.89, 42.52,
50.26, 47.10,
55.96, 50.60,
63.41, 60.11,
65.10, 65.36,
77.21, 74.25,
68.56, 67.91,
49.61, 50.25,
72.14, 70.64,
40.26, 43.29,
70.78,
75.50
```

四、计算结果输出

本程序输出文件的约定名为D1—1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、一元线性回归分析主程序

```
PROGRAM C9001
COMMON XY(500,2),X(51),Y(26),YY0(26),YY1(26),YY2(26)
```

```
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'一元线性回归分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名D1-1.DAT]:'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D1-1.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名[约定名D1-1.WRI]:'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D1-1.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入回归分析的原始数据:'
      READ(1,*,ERR=2)N,NN
      READ(1,*,ERR=2)((XY(I,J),J=1,2),I=1, N)
      WRITE(*,*)'开始进行回归分析计算:'
      WRITE(2, *)'        *****************'
      WRITE(2, *)'        *               *'
      WRITE(2, *)'        *  一元线性回归分析计算结果  *'
      WRITE(2, *)'        *               *'
      WRITE(2, *)'        *****************'
      WRITE(2, 101)N, NN
  101 FORMAT(//5X,'已知样品数:',I3/5X,'未知样品数:',I3)
      WRITE(2,'(//A)')'        原始数据表'
      WRITE(2, 102)(I,(XY(I,J),J=1, 2), I=1, N)
  102 FORMAT(/5X,'序号', 8X,'X值', 8X,'Y值'/5X,'已知样品:'/
     -(6X, I3, 2F12.3))
      IF(NN.NE.0)THEN
      N1=N+1
      NN1=N+NN
      READ(1,*)(XY(I, 1), I=N1, NN1)
      WRITE(2, 103)(I, XY(I, 1), I=N1, NN1)
  103 FORMAT(5X,'未知样品:'/(6X, I3, F12.3))
      ENDIF
      CALL YYXX(N, NN)
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X, A\)')'程序运行完闭!还继续进行计算吗?[Y/N]:'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
```

```
      STOP
    2 WRITE(*, *)'您的数据文件有错！'
      STOP
      END
```

（1）一元线性回归子程序

```
      SUBROUTINE YYXX(N, NN)
      REAL LXX, LYY, LXY
      COMMON XY(500, 2), X(51),Y(26), YY0(26), YY1(26), YY2(26)
      SX=0.
      SY=0.
      OXY=0.
      SX2=0.
      SY2=0.
      DO 1 I=1, N
      SX=SX+XY(I, 1)
      SY=SY+XY(I, 2)
      SXY=SXY+XY(I, 1)*XY(I, 2)
      SX2=SX2+XY(I, 1)*XY(I, 1)
    1 SY2=SY2+XY(I, 2)*XY(I, 2)
      X1=SX/N
      Y1=SY/N
      X2=SX*SX/N
      Y2=SY*SY/N
      XY2=SX*SY/N
      LXX=SX2-X2
      LYY=SY2-Y2
      LXY=SXY-XY2
      B=LXY/LXX
      A=Y1-B*X1
      R=LXY/SQRT(LXX*LYY)
      S=SQRT(((1-R*R)*LYY)/(N-2))
      A1=A-2*S
      A2=A+2*S
      WRITE(2, 100)A, B, A, B, R, A1, B, A2, B
  100 FORMAT(//5X,'回归方程的截距：',F10.3/5X,'回归方程的斜率：',F10.3/
     -5X,'回归方程式：Y=',F10.3,'+',F10.3,'X'/5X,'相关系数：',F10.3/
     -5X,'预测上限方程式：Y1=', F10.3, '+', F10.3, 'X'/
     -5X, '预测下限方程式：Y2=', F10.3, '+',F10.3,'X')
      IF(NN.NE.0)THEN
```

```
      N1=N+1
      NN1=N+NN
      DO 2 I=N1, NN1
    2 XY(I, 2)=A+B*XY(I, 1)
      WRITE(2,'(//A)')'           未知样品预测结果表'
      WRITE(2, 101)(I, (XY(I, J), J=1,2),I=N1, NN1)
  101 FORMAT(/5X,'序号', 8X,'X值', 8X,'Y值'/(6X, I3, 2F12.3))
      ENDIF
      CALL MAP(N, NN, A, B, A1, A2)
      END
```

(2)宽行输出打图子程序

```
      SUBROUTINE MAP(N, NN, A, B, A1, A2)
      COMMON XY(500, 2), X(51), Y(26), YY0(26),YY1(26),YY2(26)
      CHARACTER W(50), AA, BB, CC, DD, EE, FF, GG
      DATA AA, BB, CC,DD,EE,FF,GG/' ','*','I','+','.','0','Y'/
      NN1=N+NN
      XMAX=XY(1, 1)
      XMIN=XY(1, 1)
      YMAX=XY(1, 2)
      YMIN=XY(1, 2)
      DO 1 I=2, NN1
      IF(XY(I, 1).GT.XMAX)XMAX=XY(I, 1)
      IF(XY(I, 1).LT.XMIN)XMIN=XY(I, 1)
      IF(XY(I, 2).GT.YMAX)YMAX=XY(I, 2)
      IF(XY(I, 2).LT.YMIN)YMIN=XY(I, 2)
    1 CONTINUE
      DX=(XMAX-XMIN)/50
      DY=(YMAX-YMIN)/25
      XMAX=XMAX+DX*7.5
      XMIN=XMIN-DX*7.5
      YMAX=YMAX+DY*7.5
      YMIN=YMIN-DY*7.5
      DX=(XMAX-XMIN)/50
      DY=(YMAX-YMIN)/25
      DO 2 I=1, 51
    2 X(I)=XMIN+DX*(I-1)
      DO 3 I=1, 26
    3 Y(I)=YMAX-DY*(I-1)
      WRITE(2, 100)
```

```
100 FORMAT(///11X, 1HY, 15X,'一元线性回归图')
    DO 4 I=1, 26
    YY0(I)=(Y(I)-A)/B
    YY1(I)=(Y(I)-A1)/B
  4 YY2(I)=(Y(I)-A2)/B
    DO 20 I=1, 25
    W(1)=CC
    DO 5 J=2, 50
  5 W(J)=AA
    DO 7 J=1, 50
    IF(YY0(I).GT.X(J).AND.YY0(I).LE.X(J+1))GO TO 6
    GO TO 7
  6 W(J)=BB
    GO TO 8
  7 CONTINUE
  8 DO 10 J=1, 50
    IF(YY1(I).GT.X(J).AND.YY1(I).LE.X(J+1))GO TO 9
    GO TO 10
  9 W(J)=EE
    GO TO 11
 10 CONTINUE
 11 DO 13 J=1, 50
    IF(YY2(I).GT.X(J).AND.YY2(I).LE.X(J+1))GO TO 12
    GO TO 13
 12 W(J)=EE
    GO TO 14
 13 CONTINUE
 14 DO 18 J=1, NN1
    IF(XY(J, 2).LT.Y(I).AND.XY(J, 2).GE.Y(I+1))GO TO 15
    GO TO 18
 15 DO 17 K=1, 50
    IF(XY(J, 1).GT.X(K).AND.XY(J, 1).LE.X(K+1))GO TO 16
    GO TO 17
 16 IF(J.LE.N)THEN
    W(K)=FF
    ELSE
    W(K)=GG
    ENDIF
    GO TO 18
```

```
   17 CONTINUE
   18 CONTINUE
      IF(MOD(I, 5).EQ.1)GO TO 19
      WRITE(2, 101)(W(J), J=1, 50)
      GO TO 20
   19 W(1)=DD
      WRITE(2, 102)Y(I), (W(J), J=1, 50)
   20 CONTINUE
      WRITE(2, 103)Y(26), (X(J), J=1, 51, 10)
  101 FORMAT(11X, 51A1)
  102 FORMAT(1X, F10.2, 51A1)
  103 FORMAT(1X, F10.2,1H+,5(10H---------+),1X,1HX/4X,6F10.2)
      WRITE(2,'(/10X,A)')'注：(1)图中由符号*组成的直线为回归方程；'
      WRITE(2,'(15X,A)')'(2)由符号.组成的直线为预测上，下限方程；'
      WRITE(2,'(15X,A)')'(3)符号0表示已知样品点；'
      WRITE(2,'(15X,A)')'(4)符号Y表示未知样品点'
      END
```

六、一元线性回归分析计算结果

已知样品数：1

未知样品数：2

原始数据表

序号	X值	Y值
已知样品：		
1	40.250	40.160
2	46.890	42.520
3	50.260	47.100
4	55.960	50.600
5	63.410	60.110
6	65.100	65.360
7	77.210	74.250
8	68.560	67.910
9	49.610	50.250
10	72.140	70.640
11	40.260	43.000
未知样品：		
12	70.780	
13	75.500	

回归方程的截距：　　1.276

回归方程的斜率：　　.950

回归方程式：Y＝　　1.276＋　　.950X

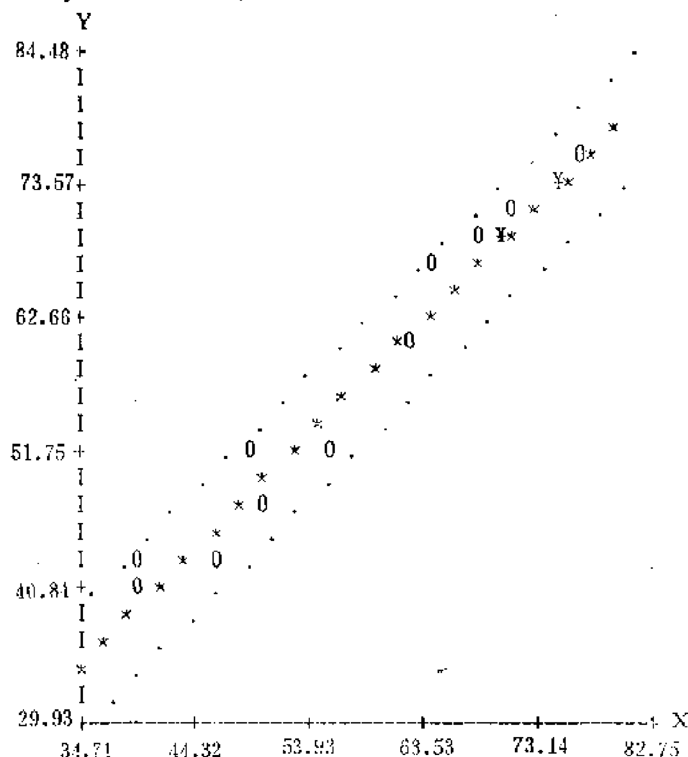
相关系数：　　.982

预测上限方程式：Y1＝　　−3.682＋　　.950X

预测下限方程式：Y2＝　　6.233＋　　.950X

未知样品预测结果表

序号	X值	Y值
12	70.780	68.483
13	75.500	72.965



附图6　一元线性回归图

注：（1）图中由符号＊组成的直线为回归方程；

（2）由符号·组成的直线为预测上、下限方程；

（3）符号0表示已知样品点；

（4）符号¥表示未知样品点

程序七　逐步回归分析

一、程序主要功能

逐步回归分析是在多元线性回归分析基础上产生的一种技巧性算法，这种算法的优点是可以从多个自变量中，自动筛选出与因变量关系最密切的自变量进入回归方程。在计算过程中根据m个自变量X1，X2，…，Xm对因变量Y的重要性逐个引进方程中，同时要对引进的变量逐个检验，通过检验保留有用的变量，剔出无用的变量，直到既不能引进也不能剔出自变量为止。

（1）本程序可以根据已知的n组观测数据

$$(Yk,\ Xik)\quad(k=1,\ 2,\ \cdots,\ n;\ i=1,\ 2,\ \cdots,\ m)$$

建立因变量与自变量之间的多元线性回归方程；所谓逐步就是指将重要的变量逐个引进回归方程。根据检验水平，如果有p个重要变量，则可以建立如下回归方程：

$$Y=b0+b1X1+b2X2+\cdots+bpXp$$

（2）引进回归方程中的变量个数决定于F检验的临界值。本程序对变量的检验可按两种方式进行，一种是由样品数m与变量数n以及用户选用的置信水平来确定临界值FF；另一种是由用户直接给定临界值FF。

（3）本程序的计算结果存入输出文件中，可由宽行打印机输出，其中包括：原始数据表，多元线性回归方程式计算结果等。

（4）对于新的未知样品，可将Xik(k=n+1, n+2, …, i=1, 2, …, p)代入回归方程：　　Y=b0+b1X1+b2X2+…+bpXp求出Yk的估计值。

二、程序符号说明

N——————————已知样品数；

NN—————————预测样品数；

M——————————变量数，其中有（M－1）个自变量，1个因变量；

M1—————————自变量数，M1=M－1；

X(500, 50)——— 原始数据增广矩阵，矩阵的行号为样品编号，列号为变量编号，自变量放在第1至第M－1列，因变量要放在最后的第M列；

XC(50)———————变量的平均值；

SR(50, 50)————离差矩阵；

R(50, 50)———— 相关系数矩阵；

KL(50)———————变量引进的标志符号；

KL(J)=0时，表示第J个变量未引进方程中；

KL(J)=1时，表示第J个变量已引进方程中；

V(50)————————变量贡献；

B(50)————————回归方程中引进的自变量待定系数；

FILENA*20—— 文件字符；

FF—————————引进或剔出变量的F临界值；

KK—————————F检验的方式；

KK=0时，由用户给定F检验置信水平（置信水平有0.1,0.5,0.01三种）并由n与m确定FF值；

KK=1时，由用户直接给定FF值；

F(34)—————— F检验表；

ALFA———————F检验置信水平。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D1－2.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
N,NN,M,KK
[FF],[ALFA]
((X(I,J),J=1,M),I=1,N)
[((X(I,J),J=1,M1),I=N1,NS)]
注：方括号[ ]中的数据为选择项
```

例如，下面的数据文件（DI-3.DAT）就是一个供用户检测本程序的数据文件：

```
32,1,5,1
2.5
13.0,7.0,26.0,19.0,11.5,
15.0,11.0,40.0,34.0,19.8,
21.0,8.0,29.0,17.0,13.7,
19.0,12.0,15.0,33.0,21.6,
27.0,11.0,13.0,27.0,22.3,
32.0,10.0,21.0,15.0.19.1,
17.0,8.0,18.0,16.0,11.7,
26.0,10.0,35.0,23.0,19.4,
14.0,6.0,14.0,18.0,10.6,
28.0,13.0,21.0,34.0,25.5,
19.0,9.0,13.0,29.0,18.7,
12.0,10.0,19.0,38.0,19.3,
23.0,8.0,25.0,17.0,15.6,
28.0,11.0,33.0,32.0,24.7,
21.0,9.0,18.0,19.0,15.3,
35.0,14.0,24.0,34.0,29.8,
16.0,6.0,19.0,14.0,10.2,
24.0,10.0,32.0,26.0,19.8,
22.0,11.0,39.0,38.0,25.3,
10.0,7.0,17.0,20.0,9.7,
18.0,8.0,34.0,22.0,14.8,
29.0,11.0,28.0,21.0,20.7,
18.0,11.0,16.0,32.0,19.6,
16.0,10.0,15.0,34.0,20.3,
18.0,7.0,23.0,14.0,11.1,
23.0,11.0,29.0,29.0,20.7,
25.0,13.0,41.0,40.0,28.9,
32.0,9.0,12.0,15.0,18.3,
36.0,11.0,37.0,18.0,21.6,
```

```
31.0,9.0,25.0,14.0,17.7,
29.0,13.0,14.0,38.0,28.3,
18.0,10.0,11.0,35.0,21.6,
18.0,10.0,11.0,35.0
```

四、计算结果输出

本程序输出文件的约定名为D1-2.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、逐步回归分析源程序

```
      PROGRAM C9002
      COMMON X(500,50),X0(50),SR(50,50),R(50,50),KL(50),V(50),B(50),
     -F(34)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'多元线性回归分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名 [约定名 D1-2.DAT]:'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D1-2.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名 [约定名 D1-2.WRI]:'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D1-2.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'正在进行计算，请等待！'
      WRITE(2,*)'         ******************'
      WRITE(2,*)'         *                *'
      WRITE(2,*)'         *  多元线性回归分析计算结果  *'
      WRITE(2,*)'         *                *'
      WRITE(2,*)'         ******************'
      READ(1,*,ERR=2)N,NN,M,KK
      IF(KK.EQ.0)THEN
      READ(1,*,ERR=2)ALFA
      IF(ALFA.EQ.0.1) OPEN(3,FILE='F10.DAT')
      IF(ALFA.EQ.0.05)OPEN(3,FILE='F05.DAT')
      IF(ALEA.EQ.0.01)OPEN(3,FILE='F01.DAT')
      READ(3,*,ERR=2)(F(I),I=1,34)
C     下面的NF为F检验的自由度.
      NF=N-M-1
      IF(NF.GT.30.AND.NF.LE.40) NF=31
```

```
      IF(NF.GT.40.AND.NF.LE.60) NF=32
      IF(NF.GT.60.AND.NF.LE.120)NF=33
      IF(NF.GT.120)             NF=34
      FF=F(NF)
      WRITE(2,101)ALFA,FF
101   FORMAT(//5X,21(1H*)/5X,'*',19(1H ),'*'/5X,'*  置信
     -水平=', F6.2,'  *'/5X,'*  F 检验值=',F6.2,'  *'/5X,'*',
      19(1H ),'*'/5X, 21(1H*))
      ELSE
      READ(1,*,ERR=2)FF
      WRITE(2,102)FF
102   FORMAT(//5X,21(1H*)/5X,'*'19(1H ),'*'/5X,'* F检验值
     -=',F6.2,'  *'/5X,'*',19(1H ),'*'/5X,21(1H*))
      ENDIF
      CALL ZHUBU(N,NN,M,FF)
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完毕！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    2 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

(1)多元线性逐步回归计算子程序

```
      SUBROUTINE ZHUBU(N,NN,M,FF)
      COMMON X(500,50),XC(50),SR(50,50),R(50,50),KL(50),V(50),
     -B(50),F(34)
C     读入已知样品的原始数据.
      READ(1,*,ERR=25)((X(I,J),J=1,M),I=1,N)
C     下面的NS为样品总数，N1为预测样品起始编号，M1为自变量总数.
      NS=N+NN
      N1=N+1
      M1=M-1
C     读入预测样品的原始数据.
      READ(1,*,ERR=25)((X(I,J),J=1,M1),I=N1,NS)
      WRITE(2,100)N,NN,M
100   FORMAT(//5X,'已知样品数：',I3/5X,'预测样品数：',I3/5X,'变量
```

```
      一数：',I3)
      WRITE(2,'(//)')
      WRITE(2,'(10X,A)')'原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
101   FORMAT(/5X,A,10(4X,A,I2,1X))
      WRITE(2,*)'     已知样品：'
      DO 1 I=1,N
1     WRITE(2,102)I,(X(I,J),J=1,M)
102   FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      IF(NN.EQ.0)GO TO 3
      WRITE(2,*)'     预测样品：'
      DO 2 I=N1,NS
2     WRITE(2,102)I,(X(I,J),J=1,M1)
C     构造相关系数矩阵R(M,M)
3     DO 5 J=1,M
      XC(J)=0.0
      DO 4 I=1,N
4     XC(J)=XC(J)+X(I,J)
5     XC(J)=XC(J)/N
      DO 6 I=1,M
      DO 6 J=1,M
      SR(I,J)=0.0
      DO 6 K=1,N
6     SR(I,J)=SR(I,J)+(X(K,I)-XC(I))*(X(K,J)-XC(J))
      DO 7 I=1,M
      DO 7 J=1,M
7     R(I,J)=SR(I,J)/(SQRT(SR(I,I))*SQRT(SR(J,J)))
C     下面的L为引进变量的总数；LL为引进与剔出变量累计次数的计数单元；
C     KL(I)为引进变量标志符号；并对L,LL,KL(I)初始化充零
      L=0
      LL=0
      DO 8 I=1,M
8     KL(I)=0
C     当：L=M1 时，表示全部自变量都已引进方程中，并且转入回归方程的系数计
C     算；否则进行引进变量计算.
1000  IF(L.EQ.M1)GO TO 3000
C     计算引进变量的方差贡献 FVMAX.
```

```
      VMAX=-10E10
      DO 10 I=1,M1
      V(I)=R(I,M)*R(I,M)/R(I,I)
      IF(KL(I).EQ.1)GO TO 10
      IF(V(I).GT.VMAX)GO TO 9
      GO TO 10
    9 VMAX=V(I)
      KI=I
   10 CONTINUE
      FVMAX=(N-L-2)*VMAX/(R(M,M)-VMAX)
C     当FVMAX>FF时，将第 I 个变量引进方程中；否则转入回归方程的系数计算.
      IF(FVMAX.GT.FF)GO TO 11
      GO TO 3000
   11 KL(KI)=1
      CALL VARY(M,KI)
      LL=LL+1
      L=0
      DO 12 I=1,M1
   12 L=L+KL(I)
C     开始引进的变量个数小于等于2，则继续引进变量；否则转入剔出变量计算.
      IF(LL.LE.1)GO TO 1000
C     当：L=0时，表示全部自变量都不能引进方程中；否则进行剔出变量计算.
 2000 IF(L.EQ.0)GO TO 4000
C     计算剔出变量的方差贡献 FVMIN.
      VMIN=10E10
      DO 14 I=1,M1
      V(I)=R(I,M)*R(I,M)/R(I,I)
      IF(KL(I).EQ.0)GO TO 14
      IF(V(I).LT.VMIN)GO TO 13
      GO TO 14
   13 VMIN=V(I)
      KI=I
   14 CONTINUE
      FVMIN=(N-L-1)*VMIN/R(M,M)
C     当：FVMIN<FF时，将第I个变量从方程中剔出；否则进行引进变量计算.
      IF(FVMIN.LT.FF)GO TO 15
      GO TO 1000
   15 KL(KI)=0
      CALL VARY(M,KI)
```

```
      LL=LL+1
      L=0
      DO 16 I=1,M1
   16 L=L+KL(I)
      GO TO 2000
C     当：L=0时，表示全部自变量都不能引进方程中；否则转入回归方程的系数计算.
 3000 IF(L.EQ.0)GO TO 4000
C     计算回归方程的待定系数 B(I)及 B0.
      DO 20 J=1,M
      V(J)=0.0
      IF(J.EQ.M)GO TO 18
      IF(KL(J).EQ.0)GO TO 20
      DO 17 I=1,N
   17 V(J)=V(J)+(X(I,J)-XC(J))*(X(I,J)-XC(J))
      GO TO 20
   18 DO 19 I=1,N
   19 V(J)=V(J)+(X(I,J)-XC(J))*(X(I,J)-XC(J))
   20 CONTINUE
      S=0.0
      DO 21 I=1,M1
      IF(KL(I).EQ.0)GO TO 21
      YI=V(M)/V(I)
      B(I)=R(I,M)*SQRT(YI)
      S=S+B(I)*XC(I)
   21 CONTINUE
      B0=XC(M)-S
C     输出多元线性回归方程式.
      WRITE(2,'(//)')
      WRITE(2,*)'     多元线性回归方程式'
      WRITE(2,103)B0
  103 FORMAT(/5X,25(1H*)/5X,1H*,23(1H ),1H*/5X,5H*      Y=,
     -F12.4,7(1H ),1H*)
      DO 22 I=1,M1
      IF(KL(I).EQ.0)GO TO 22
      WRITE(2,104)B(I),I
  104 FORMAT(5X,5H*      +,F12.4,2HX(,I2,4H)   *)
   22 CONTINUE
      WRITE(2,105)
  105 FORMAT(5X,1H*,23(1H ),1H*/5X,25(1H*))
```

```
C       计算并输出复相关系数RY及剩余标准差SY.
        RY=SQRT(1-R(M,M))
        SY=SQRT(V(M))*SQRT(R(M,M)/(N-L-1))
        WRITE(2,106)RY,SY
  106   FORMAT(//5X,'复相关系数:',F12.4/5X,'剩余标准差:',F12.4)
C       如果有预测样品，则计算并输出预测值Y.
        IF(NN.EQ.0)GO TO 4000
        DO 23 I=N1,NS
        X(I,M)=B0
        DO 23 J=1,M1
        IF(KL(J).EQ.0)GO TO 23
        X(I,M)=X(I,M)+B(I)*X(I,J)
   23   CONTINUE
        WRITE(2,'(//)')
        WRITE(2,'(10X,A)')'预测样品数据表'
        MXY=M
        IF(M.GT.10)MXY=10
        WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
        DO 24 I=N1,NS
   24   WRITE(2,102)I,(X(I,J),J=1,M)
C       返回主程序.
 4000   RETURN
   25   WRITE(*,*)'您的数据文件有错!'
        END
```

(2)引进或剔出变量的矩阵变换子程序

```
        SUBROUTINE VARY(M,KI)
        COMMON X(500,50),XC(50),SR(50,50),R(50,50),KL(50),V(50),B(50),
       -F(34)
        DO 6 I=1,M
        W=R(KI,KI)
        IF(I.EQ.KI)GO TO 3
        DO 2 J=1,M
        IF(J.EQ.KI)GO TO 1
        SR(I,J)=R(I,J)-R(I,KI)*R(KI,J)/W
        GO TO 2
    1   SR(I,J)=-R(I,KI)/W
    2   CONTINUE
        GO TO 6
```

```
  3  DO 5 J=1,M
   IF(J.EQ.KI)GO TO 4
   SR(I,J)=R(KI,J)/W
   GO TO 5
4  SR(I,J)=1/W
5  CONTINUE
6  CONTINUE
   DO 7 I=1,M
   DO 7 J=1,M
7  R(I,J)=SR(I,J)
   END
```

六、多元线性回归分析计算结果

置信水平= .10

F检验值 =2.91

已知样品数：32

预测样品数： 1

变量数：5

原始数据表

样品序号	变量1	变量2	变量3	变量4	变量5
已知样品：					
1	13.000	7.000	26.000	19.000	11.500
2	15.000	11.000	40.000	34.000	19.800
3	21.000	8.000	29.000	17.000	13.700
4	19.000	12.000	15.000	33.000	21.600
5	27.000	11.000	13.000	27.000	22.300
6	32.000	10.000	21.000	15.000	19.100
7	17.000	8.000	18.000	16.000	11.700
8	26.000	10.000	35.000	23.000	19.400
9	14.000	6.000	14.000	18.000	10.600
10	28.000	13.000	21.000	34.000	25.500
11	19.000	9.000	13.000	29.000	18.700
12	12.000	10.000	19.000	38.000	19.300
13	23.000	8.000	25.000	17.000	15.600
14	28.000	11.000	33.000	32.000	24.700
15	21.000	9.000	18.000	19.000	15.300
16	35.000	14.000	24.000	34.000	29.800
17	16.000	6.000	19.000	14.000	10.200
18	24.000	10.000	32.000	26.000	19.800
19	22.000	11.000	39.000	38.000	25.300

20	10.000	7.000	17.000	20.000	9.700
21	18.000	8.000	34.000	22.000	14.800
22	29.000	11.000	28.000	21.000	20.700
23	18.000	11.000	16.000	32.000	19.600
24	16.000	10.000	15.000	34.000	20.300
25	18.000	7.000	23.000	14.000	11.100
26	23.000	11.000	29.000	29.000	20.700
27	25.000	13.000	41.000	40.000	28.900
28	32.000	9.000	12.000	15.000	18.300
29	36.000	11.000	37.000	18.000	21.500
30	31.000	9.000	25.000	14.000	17.700
31	29.000	13.000	14.000	38.000	28.300
32	18.000	10.000	11.000	35.000	21.600

预测样品：

33	18.000	10.000	11.000	35.000

多元线性回归方程式

```
* * * * * * * * * * * * *
*                       *
*   Y=-5.9318           *
*       +2.5380X(2)     *
*                       *
* * * * * * * * * * * * *
```

复相关系数： .9566

剩余标准差： 1.6267

预测样品数据表

样品序号	变量1	变量2	变量3	变量4	变量5
33	18.000	10.000	11.000	35.000	19.448

七、F检验时的三个数据文件

下面是逐步回归分析进行F检验时的三个数据文件，需要存入当前工作盘。

（1）置信水平为0.10时的F10.DAT文件：

39.86, 8.53, 5.54, 4.54, 4.06, 3.78, 3.59, 3.46, 3.36, 3.29, 3.23, 3.18, 3.14, 3.10, 3.07, 3.05, 3.03, 3.01, 2.99, 2.97, 2.96, 2.95, 2.94, 2.93, 2.92, 2.91, 2.90, 2.89, 2.89, 2.88, 2.84, 2.79, 2.75, 2.71

（2）置信水平为0.05时的F05.DAT文件：

161.4, 18.51, 10.13, 7.71, 6.61, 5.99, 5.59, 5.32, 5.12, 4.96, 4.84, 4.75, 4.67, 4.60, 4.54, 4.49, 4.45, 4.41, 4.38, 4.35, 4.32, 4.30, 4.28, 4.26, 4.24, 4.23, 4.21, 4.20, 4.18, 4.17, 4.08, 4.00, 3.92, 3.84

（3）置信水平为0.01时的F01.DAT文件：

4052.0, 98.50, 34.12, 21.20, 16.26, 13.75, 12.25, 11.26, 10.56, 10.04, 9.65,

9.33, 9.07, 8.86, 8.68, 8.53, 8.40, 8.29, 8.18, 8.10, 8.02, 7.95, 7.88, 7.82, 7.77, 7.72, 7.68, 7.64, 7.60, 7.56, 7.31, 7.08, 6.85, 6.63

程序八　多项式趋势分析

一、程序主要功能

趋势分析的基本原理是认为任何一个地质空间曲面都可以分解为三个部分，即：

地质空间曲面＝区域性因素＋局部性因素＋随机性因素
（综合值）（背景值）（异常值）（干扰值）

如果能以多项式数学曲面（趋势面）代替区域性因素（背景值），并且用统计方法消除随机性因素（干扰值），则能达到突出局部性因素（异常值）之目的。由于随机性因素服从正态分布，那么可以用

地质空间曲面－趋势面

表示异常值。因而，应该设法使趋势面尽量接近背景值。

（1）本程序可以根据原始数据，构造出1次至10次多项式趋势面供用户选择使用。为了保证精度，计算高次趋势面时，数组X，Y可采用双精度。

（2）本程序可以由宽行打印机输出：原始数据点位置图（MAP0），观测值平面插值图（MAP1），趋势图（MAP2），残差图（MAP3）。其中：

原始数据点位置图：表示数据点之间的相对位置；

观测值平面插值图：是按“距离倒数平方加权法”进行计算的，用户可以任选其中的全点法，近点法，圆内法，或象限法；

趋势图：是由趋势面方程计算的，本程序可以连续计算1次至10次多项式趋势方程并打印出趋势面图，用户可指定趋势分析的最低次数KMIN及最高次数KMAX；

残差图：残差值等于（观测值-趋势值）。残差图是由计算观测值平面插值图时，形成的数据文件D2-1.Z与计算趋势图时，形成的数据文件D2-1.ZN之间的对应点数值相减得到。

（3）本程序还可以进行复合趋势分析计算，即：可以对残差值再进行趋势分析，此时可令数据文件中的KN＝1，并且要指定残差趋势分析的最低次数KMIN1及最高次数KMAX1．需要指出，残差趋势分析的最低次数必须大于趋势分析的最高次数，即：KMIN1＞KMAX．

二、程序符号说明

N———————原始数据点数（N小于等于600）；

MP——————原图比例尺（例如：5万，20万等）；

XS ——————原图横向（X轴方向）长度（单位为厘米）；

YS ——————原图纵向（Y轴方向）长度（单位为厘米）；

XM——————宽行打印机100列字符的长度（单位为厘米）；

（注：宽行打印机100列字符的标准长度为25.4厘米）

MY——————宽行打印机在100列字符长度内的打印行数；

（注：宽行打印机100列字符长度内的标准打印行数为60行）

MN——————平面图的接图次数，例如、MN=2时，则图形的边长增大一倍（面积增大四倍），输出图件可由两幅宽行打印图拼接而成。

KM——————平面图的分带数（KM小于等于10），例如：KM=5时，相当于将研究区分为：好，较好，中等，较差，差，五个等级。

KK ——————平面插值（距离倒数平方加权法）的计算方式；

KK=1时，为全点法；

KK=2时，为近点法（最近的5个点）；

KK=3时，为圆内法（圆的面积为原图面积的十分之一）；

KK=4时，为象限法（直角四象限中各象限取一个近点）；

KMIN ———趋势分析的最低次数；

KMAX ———趋势分析的最高次数；

KN——————是否作残差趋势分析的标志符号；

KN=0时，为不作；

KN=1时，为要作；

KL——————原始数据点的坐标（X，Y）选用方式；

KL=0时，为相对坐标；

KL=1时，为绝对坐标；

$KMIN_1$ ———残差趋势分析的最低次数；

$KMAX_1$ ——残差趋势分析的最高次数；

X(500) ———原始数据点的横坐标（单位为厘米）；

Y(500) ———原始数据点的纵坐标（单位为厘米）；

原始数据点的横坐标X及纵坐标Y可按下面两种方法确定：

（1）相对坐标（KL=0）：是以输出图形的左下图角为坐标原点（0，0），每个原始数据点的相对坐标（X，Y），是以厘米为单位由图上直接测量。

（2）绝对坐标（KL=1）：直接采用野外的实际测量数据（单位为米），并且要给出输出图形左下角的坐标值（X_0，Y_0）。

Z(500) ———原始数据点的观测值（单位由用户确定）；

H（11）———平面图的分带区间；

ZN(500) ——趋势值；

ZQ(500) ——残差值；

AZ(66,67) —趋势方程系数矩阵（方程的最大次数为10次）；

III(66) ———趋势方程的X项系数方次；

JJJ(66)———趋势方程的Y项系数方次；

W(100)———打印字符工作单元；

FILENA ——文件名字符；

NOYES$ ———是否继续进行计算的标志符。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D2-1.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
MP, XS, YS, KL, KK, MN, KM, KN, XM, MY
[X0, Y0]
N
(X(J), Y(J), J=1, N)
(Z(J), J=1, N)
KMIN, KMAX
[KMIN1, KMAX1]
```

注：方括号[ ]中的数据为选择项

例如，下面的数据文件（D2-1.DAT）就是一个供用户检测本程序的数据文件：

```
100000, 6.4, 6.4, 0, 1, 1, 10, 0, 25.4, 60
52
0.3, 6.1, 1.4, 6.2, 2.4, 6.1, 3.6, 6.2,
5.7, 6.2, 1.6, 5.2, 2.9, 5.1, 3.4, 5.3,
3.4, 5.7, 4.8, 5.6, 5.3, 5.0, 6.2, 5.2,
0.2, 4.3, 0.9, 4.2, 2.3, 4.8, 2.5, 4.5,
3.0, 4.5, 3.5, 4.5, 4.1, 4.6, 4.9, 4.2,
6.3, 4.3, 0.9, 3.2, 1.7, 3.8, 2.4, 3.8,
3.7, 3.5, 4.5, 3.2, 5.2, 3.2, 6.3, 3.4,
0.3, 2.4, 2.0, 2.7, 3.8, 2.3, 6.3, 2.2,
0.6, 1.7, 1.5, 1.8, 2.1, 1.8, 2.1, 1.1,
3.1, 1.1, 4.5, 1.8, 5.5, 1.7, 5.7, 1.0,
6.2, 1.0, 0.4, 0.5, 1.4, 0.6, 1.4, 0.1,
2.1, 0.7, 2.3, 0.3, 3.1, 0.01, 4.1, 0.8,
5.4, 0.4, 6.0, 0.1, 5.7, 3.0, 3.6, 6.0,
870.0, 793.0, 755.0, 690.0, 800.0, 800.0, 730.0, 728.0,
710.0, 780.0, 804.0, 855.0, 830.0, 813.0, 762.0, 765.0,
740.0, 765.0, 760.0, 790.0, 820.0, 855.0, 812.0, 773.0,
812.0, 827.0, 805.0, 840.0, 890.0, 820.0, 873.0, 875.0,
873.0, 865.0, 847.0, 862.0, 908.0, 855.0, 850.0, 882.0,
910.0, 940.0, 915.0, 890.0, 880.0, 870.0, 880.0, 960.0,
890.0, 860.0, 830.0, 705.0
```

1, 1

四、计算结果输出

本程序输出文件的约定名为D2-1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、多项式趋势分析主程序

```
      PROGRAM C9004
      COMMON X(500), Y(500), Z(500), H(11),ZN(500),ZQ(500), AZ(66,67),
     —III(66), JJJ(66)
      CHARACTER FILENA *20, NOYES
    1 WRITE(*, '(1X, 30(1H )\)')
      WRITE(*, '(1X, A)')'多项式趋势分析'
      WRITE(*, '(1X, A\)')'请输入您的数据文件名[约定名D2-1.DAT]: '
      READ(*, '(A)')FILENA
      IF(FILENA.EQ.'    ')FILENA='D2-1.DAT'
      OPEN(1, FILE=FILENA)
      WRITE(*, '(1X, A\)')'请输入您的输出文件名[约定名D2-1.WRI]: '
      READ(*, '(A)')FILENA
      IF(FILENA.EQ.'    ')FILENA='D2-1.WRI
      OPEN(2,FILE=FILENA,STATUS='NEW')
      XM=25.4
      MY=60
      RK=10.0
      JJ=5
      WRITE(*, *)'开始读入趋势分析的原始数据: '
      READ(1,*,ERR=4)MP, XS, YS, KL, KK, MN, KM, KN, XM, MY
      PM=MP*XS/(MN*XM)
      MM=INT(PM)
      KY=MY*MN*YS/XS
      R2=SQRT(XS*YS/(3.141592654*RK))
      S=YS/KY
      S2=XS/(MN*100.)
      SR=S2/10.
      DS=XS/MN
      WRITE(2, *)'        ********************'
      WRITE(2, *)'        *                  *'
      WRITE(2, *)'        *  多项式趋势分析计算结果  *'
      WRITE(2, *)'        *                  *'
      WRITE(2, *)'        ********************'
```

```
      WRITE(2,101)MP, XS, YS, MN, KM, KK, KY, KL, KN
  101 FORMAT(//5X,'原图比例尺:', I7/5X, '原图横向长度:', F6.2/
     -5X,'原图纵向长度:', F6.2/5X, '接图次数:', I2/5X,'平面图分带数:',
     -I2/5X, '平面插值计算方式:', I2/5X, '平面图的扫描行数:', I2/5X,
     -'坐标选习方式:', I2/5X, '是否作残差趋势分析:', I2)
C     打开临时性文件(D2-1.Z),准备写入"观测值平面插值".
      OPEN(3,FILE='D2-1.Z',STATUS='NEW')
C     打开临时性文件(D2-1.ZN),准备写入"趋势值".
      OPEN(4,FILE='D2-1.ZN',STATUS='NEW')
      IF(KL.NE.0)THEN
      READ(1,*,ERR=4)X0,Y0
      WRITE(2,102)X0,Y0
  102 FORMAT(//5X,'绝对坐标的X0:', F12.3/5X, '绝对坐标的Y0:',F12.3)
      ENDIF
      READ(1, *, ERR=4)N
      READ(1, *, ERR=4)(X(J), Y(J), J=1, N)
      READ(1, *, ERR=4)(Z(J), J=1, N)
      READ(1, *, ERR=4)KMIN, KMAX
      WRITE(*, *)'开始进行趋势分析计算:'
      WRITE(2,103)KMIN, KMAX
  103 FORMAT(//5X,'趋势分析最低次数:', I2/5X, '趋势分析最高次数:', I2)
      WRITE(2, 104)(J, X(J), Y(J), Z(J), J=1,N)
  104 FORMAT(//'                    趋势分析原始数据表'/11X,'序号', 6X,
     -'横坐标', 6X, '纵坐标',6X,'观测值'/(11X, I5, 1X, 3F12.3))
      IF(KL.NE.0)CALL MP1(N, MP, X0, Y0)
      CALL QS(N, KY, XS, YS, KL, MN, MM, S,S2,SR,R2,KM, KK,JJ,
     -KMIN, KMAX)
      IF(KN.EQ.0)GO TO 3
      WRITE(*, *)'开始读入残差趋势分析的原始数据:'
      D0 2 J=1, N
      Z(J)=ZQ(J)
    2 ZQ(J)=0.0
      READ(1, *, ERR=4)KMIN1, KMAX1
      WRITE(2, 105)KMIN1, KMAX1
  105 FORMAT(//5X,'残差趋势分析最低次数:',I2/5X,'残差趋势分析最高次数:',
     -I2)
      IF(KL.NE.0)CALL MP2(N, MP, X0, Y0)
      WRITE(2, 106)(J, X(J), Y(J), Z(J), J=1,N)
  106 FORMAT(//'                    残差趋势分析原始数据表'/11X,'序号',
```

```
-6X,'横坐标', 6X,'纵坐标',6X,'观测值'/(11X, I3, 1X, 3F12.3))
   WRITE(*, *)'开始进行残差趋势分析计算;'
   IF(KL.NE.0)CALL MP1(N, MP, X0, Y0)
   CALL QS(N, KY, XS, YS, KL, MN, MM, S, S2, SR, R2, KM,KK,
  -JJ, KMIN1, KMAX1)
 3 CLOSE(1)
   CLOSE(2)
   CLOSE(3,STATUS='DELETE')
   CLOSE(4,STATUS='DELETE')
   WRITE(*,'(1X,A\)')'程序运行完闭!还继续进行计算吗? [Y/N]:'
   READ(*,'(A)')NOYES
   IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y') GO TO 1
   STOP
 4 WRITE(*, *)'您的数据文件有错!'
   STOP
   END
```

(1)多项式趋势分析子程序

```
   SUBROUTINE  QS(N, KY, XS, YS, KL, MN, MM, S, S2, SR, R2,
  -KM, KK, JJ, KMIN, KMAX)
   COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
  -AZ(66,67), III(66), JJJ(66)
   WRITE(*, *)'正在计算:原始数据点位置图(MAP0);'
   DO 1000 L=1,MN
   S1=(L-1)*XS/MN
1000 CALL MAP0(N, KY, MN, MM, L, S, S1, S2)
   WRITE(*, *)'正在计算:观测值平面插值图(MAP1);'
   DO 2000 L=1,MN
   S1=(L-1)*XS/MN
2000 CALL MAP1(N, KY, MN, MM, L, S, S1, S2, SR, R2, KM, KK,JJ)
C     建立趋势方程系数矩阵(AZ).
   III(1)=0
   JJJ(1)=0
   DO 1 K=1, 10
   K0=K*(K+1)/2+1
   KI=(K+1)*(K+2)/2
   DO 1 I=K0, KI
   III(I)=KI-I
 1 JJJ(I)=I-K0
   DO 18 K=KMIN, KMAX
```

```
      WRITE(*,'(1X, A, I2, A\)')'开始计算', K,'次趋势分析; '
      KI=(K+1)*(K+2)/2
      KJ=KI+1
      WRITE(2,100)(III(I), I=1, KI)
  100 FORMAT(//5X,'趋势方程X项系数方次: '/(5X, 20I5 ) )
      WRITE(2,101)(JJJ(I), I=1, KI)
  101 FORMAT(5X,'趋势方程Y项系数方次:'/(5X, 20I5 ) )
      DO 5 I=1,KI
      DO 5 J=I, KJ
      AZ(I, J)=0.0
      IF(J.EQ.KJ)GO TO 3
      DO 2 L=1,N
    2 AZ(I,J)=AZ(I,J)+X(L)**(III(I)+III(J))*Y(L)**(JJJ(I)+JJJ(J))
      GO TO 5
    3 DO 4 L=1, N
    4 AZ(I,J)=AZ(I, J)+X(L)**III(I)*Y(L)**JJJ(I)*Z(L)
    5 CONTINUE
      DO 6 I=2, KI
      J1=I-1
      DO 6 J=1, J1
    6 AZ(I,J)=AZ(J,I)
C     计算并输出观测值的平均值(ZCP).
      ZCP=AZ(1, KJ)/N
      WRITE(2, 102)ZCP
  102 FORMAT (//5X,'观测值平均值:',F12.3 )
C     如果趋势分析的次数大于3时,因为系数矩阵的项数太多,则在宽行纸上
C     不便于输出趋势方程系数矩阵.
      IF(K.GT.3)GO TO 8
      WRITE(2, 103)K
  103 FORMAT(//5X,'趋势方程系数矩阵AZ(', I1,' ):' )
      DO 7 I=1,KI
    7 WRITE(2, 104)(AZ(I, J), J=1, KJ)
  104 FORMAT(1X, 11F12.3)
C     用高斯消元法求解趋势方程.
    8 KI1=KI-1
      EP=10.0E-10
      DO 12 L=1, KI1
      P=0.0
      DO 9 I=L, KI
```

```
      IF(ABS(AZ(I, L)).LE.ABS(P))GO TO 9
      P=AZ(I, L)
      I0=I
    9 CONTINUE
      IF(ABS(P).LE.EP)GO TO 19
      IF(I0.EQ.L)GO TO 11
      DO 10 J=L, KJ
      T=AZ(L, J)
      AZ(L, J)=AZ(I0, J)
   10 AZ(I0, J)=T
   11 P=1.0/P
      K1=L+1
      DO 12 J=K1, KJ
      AZ(L, J)=AZ(L, J)*P
      DO 12 I=K1, KI
   12 AZ(I, J)=AZ(I,J)-AZ(I, L)*AZ(L, J)
      AZ(KI, KJ)=AZ(KI, KJ)/AZ(KI, KI)
      DO 14 L=1, KI1
      I=KI-L
      K1=I+1
      P=0.0
      DO 13 J=K1, KI
   13 P=P+AZ(I, J)*AZ(J, KJ)
   14 AZ(I, KJ)=AZ(I, KJ)-P
C     趋势方程有解（令KIJ=0作为标志）；并输出趋势方程的系数.
      KIJ=0
      WRITE(*,'(1X, I2, A)')K,'次趋势方程有解; '
      WRITE(2, 105)K, KIJ
  105 FORMAT(//5X, I2,'次趋势方程有解; KIJ=',I2)
      WRITE(2, 106)
  106 FORMAT(//5X,'  趋势方程系数'/5X,30(1H*)/5X, 1H*, 28(1H  ),
     -1H*)
      DO 15 I=1, KI
      L=I-1
   15 WRITE(2, 107)L, AZ(I, KJ)
  107 FORMAT(5X, 4H*  B(, I2, 2H)=, F20.8, 2H  *)
      WRITE(2, 108)
  108 FORMAT(5X, 1H*, 28(1H  ), 1H*/5X, 30(1H*))
C     计算并输出：趋势值（ZN），残差值（ZQ），总离差平方和（SS），残差平方
```

```
C      和（V），回归平方和（U），拟合度（C），F检验值（FF）。
       V=0.0
       SS=0.0
       DO 17 L=1, N
       ZN(L)=0.0
       DO 16 I=1, KI
    16 ZN(L)=ZN(L)+AZ(I, KJ)*X(L)**III(I)*Y(L)**JJJ(I)
       ZQ(L)=Z(L)-ZN(L)
       V=V+ZQ(L)**2
    17 SS=SS+(Z(L)-ZCP)**2
       U=SS-V
       C=U/SS
       FF=(U/KI1)/(V/(N-KI1-1))
       WRITE(2, 109)(I, X(I), Y(I), Z(I), ZN(I), ZQ(I), I=1, N)
   109 FORMAT(//'                    趋势分析计算结果'/11X,
     -'序号', 6X,'横坐标', 6X,'纵坐标', 6X,'观测值', 6X,'趋势值', 6X,'残差
     -值'/(11X, I3, 1X, 5F12.3))
       IF(KL.NE.0)WRITE(2, *)'          注：上表中的横坐标（X）及纵坐标
     -（Y），已换算成本图的相对坐标.'
       WRITE(2, 110)SS, V, U, C, FF
   110 FORMAT(//5X,'总离差平方和:', F20.3/5X,'残差平方和:', F20.3/5X,
     -'回归平方和:', F20.3/5X,'拟合度:', F20.3/5X,'F检验值:', F20.3)
       WRITE(*,'(1X, A, I2, A)')'正在计算:', K,'次趋势图（MAP2）;'
       DO 3000 L=1, MN
       S1=(L-1)*XS/MN
  3000 CALL MAP2(N, KY, MN, MM, L, S, S1, S2, KI, KJ, KM)
       WRITE(*,'(1X, A, I2, A)')'正在计算:', K,'次残差图（MAP3）;'
       DO 4000 L=1, MN
       S1=(L-1)*XS/MN
  4000 CALL MAP3(N, KY, MN, MM, L, S, S1, S2, KM)
    18 CONTINUE
       RETURN
C      趋势方程无解（令KIJ=1作为标志）；返回主程序.
    19 KIJ=1
       WRITE(*,'(1X, I2, A)')K,'次趋势方程无解！'
       WRITE(2,111)K, KIJ
   111 FORMAT(//5X, I2,'次趋势方程无解！KIJ=', I3)
       END
```

（2）原始数据点位置图子程序

```
      SUBROUTINE MAP0(N, KY, MN, MM, L, S, S1, S2)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      CHARACTER*1 A, B, BI(10),W(100)
      DATA BI/'1','2','3','4','5','6','7','8','9','0'/
      DATA A,B/' ', '+'/
      WRITE(2,'(/////)')
      WRITE(2,'(56X, A)')'原始数据点位置图'
      CALL WR1(MM, MN, L)
      DO 1 I=1, 10
      N1=N/10**I
      IF(N1.EQ.0)GO TO 2
    1 CONTINUE
    2 N1=I
      DO 10 I=1, KY
      YI=(KY-I+1)*S
      YI1=YI-S
      DO 3 J=1,100
    3 W(J)=A
      DO 9 M=1, N
      IF(Y(M).GT.YI1.AND.Y(M).LE.YI)GO TO 4
      GO TO 9
    4 DO 8 J=1,100
      XJ=S1+S2*J
      XJ1=XJ+S2
      IF(X(M).GE.XJ.AND.X(M).LT.XJ1)GO TO 5
      GO TO 8
    5 W(J)=B
      M1=M
      DO 7 JJ=1,N1
      J2=N1-JJ
      M2=M1/10**J2
      IF(M2.EQ.0)GO TO 6
      W(J+JJ)=BI(M2)
      GO TO 7
    6 W(J+JJ)=BI(10)
    7 M1=M1-M2*10**J2
      GO TO 9
    8 CONTINUE
```

```
    9 CONTINUE
   10 WRITE(2, 100)(W(J), J=1,100)
  100 FORMAT(10X, 7HI()I  I, 100A1, 7HI  I()I)
      CALL WR2
      END
```

（3）观测值平面插值图子程序

```
      SUBROUTINE MAP1(N, KY, MN, MM, L, S, S1, S2, SR,R2, KM,
     -KK, JJ)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      CHARACTER*1 W(100)
      WRITE(2, '(/////)')
      WRITE(2,'(56X, A)')'观测值平面插值图'
      REWIND 3
      IF(N.LT.JJ)JJ=N
      ZMAX=Z(1)
      ZMIN=Z(1)
      DO 1 I=1, N
      IF(Z(I).GT.ZMAX)ZMAX=Z(I)
      IF(Z(I).LT.ZMIN)ZMIN=Z(I)
    1 CONTINUE
      DH=(ZMAX-ZMIN)/KM
      KM1=KM+1
      DO 2 I=1, KM1
    2 H(I)=ZMIN+DH*(I-1)
      H(I)=H(1)-DZ
      H(KM1)=H(KM1)+DZ
      CALL WR1(MM, MN, L)
      DO 4 I=1, KY
      YI=(KY-I+1)*S
      DO 3 J=1, 100
      XJ=S1+S2*J
      CALL COMP(N, KK, YI, XJ, SR, JJ, R2, ZZ)
      WRITE(3, *)ZZ
      CALL ZH(W, J, KM, ZZ)
    3 CONTINUE
    4 WRITE(2, 101)(W(J), J=1, 100)
  101 FORMAT(10X, 7HI()I  I, 100A1, 7HI  I()I)
      CALL WR2
```

```
      H(1)=H(1)+DZ
      H(KM1)=H(KM1)-DZ
      CALL WH(KM)
      END
```

（4）趋势图子程序

```
      SUBROUTINE MAP2(N, KY, MN, MM, L, S, S1, S2, KI, KJ, KM)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      CHARACTER*1 W(100)
      WRITE(2,'(/////)')
      WRITE(2,'(60X,A)')'趋势图'
      REWIND 4
      ZMAX=ZN(1)
      ZMIN=ZN(1)
      DO 1 I=1,N
      IF(ZN(I).GT.ZMAX)ZMAX=ZN(I)
      IF(ZN(I).LT.ZMIN)ZMIN=ZN(I)
    1 CONTINUE
      DZ=(ZMAX-ZMIN)/KM
      KM1=KM+1
      DO 2 I=1, KM1
    2 H(I)=ZMIN+DZ*(I-1)
      H(1)=H(1)-DZ
      H(KM1)=H(KM1)+DZ
      CALL WR1(MM, MN, L)
      DO 5 I=1, KY
      YI=(KY-I+1)*S
      DO 4 J=1, 100
      XJ=S1+S2*J
      ZZ=0.0
      DO 3 M=1,KI
    3 ZZ=ZZ+AZ(M, KJ)*XJ**III(M)*YI**JJJ(M)
      WRITE(4, *)ZZ
      CALL ZH(W, J, KM, ZZ)
    4 CONTINUE
    6 WRITE(2, 101)(W(J), J=1, 100)
  101 FORMAT(10X,7HI()I  I, 100A1, 7HI  I()I)
      CALL WR2
      H(1)=H(1)+DZ
```

```
      H(KM1)=H(KM1)-DZ
      CALL WH(KM)
      END
```

（5）残差图子程序

```
      SUBROUTINE MAP3(N, KY, MN, MM, L, S, S1, S2, KM)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66,67), III(66), JJJ(66)
      CHARACTER*1 W(100)
      WRITE(2,'(/////)')
      WRITE(2,'(60X,A)')残差图'
      REWIND 3
      REWIND 4
      ZMAX=ZQ(1)
      ZMIN=ZQ(1)
      DO 1 I=1, N
      IF(ZQ(I).GT.ZMAX)ZMAX=ZQ(I)
      IF(ZQ(I).LT.ZMIN)ZMIN=ZQ(I)
    1 CONTINUE
      DZ=(ZMAX-ZMIN)/KM
      KM1=KM+1
      DO 2 I=1,KM1
    2 H(I)=ZMIN+DZ*(I-1)
      H(1)=H(1)-DZ
      H(KM1)=H(KM1)+DZ
      CALL WR1(MM, MN, L)
      DO 4 I=1, KY
      YI=(KY-I+1)*S
      DO 3 J=1, 100
      XJ=S1+S2*J
      READ(3, *)Z1
      READ(4, *)Z2
      ZZ=Z1-Z2
      CALL ZH(W, J, KM, ZZ)
    3 CONTINUE
    4 WRITE(2, 101)(W(J),J=1, 100)
  101 FORMAT(10X, 7HI()I  I, 100A1, 7HI  I()I)
      CALL WR2
      H(1)=H(1)+DZ
      H(KM1)=H(KM1)-DZ
```

```
      CALL WH(KM)
      END
```
(6)选择字符子程序
```
      SUBROUTINE ZH(W, J, KM, ZZ)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      CHARACTER*1 AI(10), A, W(100)
      DATA AI/'A,'.','B',': ','C',',','D','; ','E','-'/
      DATA A/'  '/
      DO 1 K=1,KM
      IF(H(K).LE.ZZ.AND.H(K+1).GT.ZZ)GO TO 2
    1 CONTINUE
      W(J)=A
    2 W(J)=AI(K)
      RETURN
      END
```
(7)打印图例子程序
```
      SUBROUTINE WH(KM)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      CHARACTER*1 AI(10), W(10)
      DATA AI/'A', '.', 'B', ': ', 'C', ',','D', '; ', 'E', '-'/
      WRITE(2, '(22X, A)')'图          例'
      DO 2 I=1, KM
      DO 1 J=1,10
    1 W(J)=AI(I)
    2 WRITE(2, 100)(W(J), J=1, 10), H(I), H(I+1)
  100 FORMAT(10X, 10A1, 5X, F10.3, 5H-----, F10.3)
      END
```
(8)平面插值计算子程序
```
      SUBROUTINE COMP(N, KK, YI, XJ, SR, JJ, R2, ZZ)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      DIMENSION RI(500), KZ(500), RMIN(4), KR(4), ZMINS(4)
      GO TO(1000, 2000, 3000, 4000)KK
C     全点插值法:
 1000 S1=0.
      S2=0.
      DO 101 M=1, N
```

```
      XR=XJ-X(M)
      YR=YI-Y(M)
      R=SQRT(XR*XR+YR*YR)
      IF(R.LT.SR)GO TO 102
      S1=S1+Z(M)/(R*R)
  101 S2=S2+1/(R*R)
      ZZ=S1/S2
      GO TO 103
  102 ZZ=Z(M)
  103 RETURN
C     近点插值法：
 2000 DO 201 M=1, N
      XR=XJ-X(M)
      YR=YI-Y(M)
      RI(M)=SQRT(XR*XR+YR*YR)
      IF(RI(M).LT.SR)GO TO 206
  201 KZ(M)=M
      J=N
  202 IF(J.GT.1)THEN
      J=J/2
  203 K=0
      I1=N-J
      DO 204 I=1, I1
      IF(RI(I).GT.RI(I+J)THEN
      T=RI(I)
      RI(I)=RI(I+J)
      RI(I+J)=T
      L=KZ(I)
      KZ(I)=KZ(I+J)
      KZ(I+J)=L
      K=1
      ENDIF
  204 CONTINUE
      IF(K.EQ.1)GO TO 203
      GO TO 202
      ENDIF
      S1=0.
      S2=0.
      DO 205 M=1, JJ
```

```
      MZ=KZ(M)
      S1=S1+Z(MZ)/(RI(M)*RI(M))
  205 S2=S2+1/(RI(M)*RI(M))
      ZZ=S1/S2
      GO TO 207
  206 ZZ=Z(M)
  207 RETURN
C     圆内插值法：
 3000 S1=0.
      S2=0.
      KS=0
      DO 301 M=1, N
      XR=XJ-X(M)
      YR=YI-Y(M)
      R=SQRT(XR*XR+YR*YR)
      IF(R.LT.SR)GO TO 302
      IF(R.GT.R2)GO TO 301
      S1=S1+Z(M)/(R*R)
      S2=S2+1/(R*R)
      KS=KS+1
  301 CONTINUE
      IF(KS.EQ.0)GO TO 303
      ZZ=S1/S2
      GO TO 304
  302 ZZ=Z(M)
      GO TO 304
  303 ZZ=-10E30
  304 RETURN
C     象限插值法：
 4000 DO 401 M=1, 4
  401 KR(M)=0
      DO 402 M=1, 4
  402 RMIN(M)=10E10
      DO 411 M=1, N
      XR=XJ-X(M)
      YR=YI-Y(M)
      RI(M)=SQRT(XR*XR+YR*YR)
      IF(RI(M).LT.SR)GO TO 414
      IF(X(M).GT.XJ.AND.Y(M).GT.YI)GO TO 403
```

```
      IF(X(M).LT.XJ.AND.Y(M).GT.YI)GO TO 405
      IF(X(M).LT.XJ.AND.Y(M).LT.YI)GO TO 407
      IF(X(M).GT.XJ.AND.Y(M).LT.YI)GO TO 409
  403 IF(RI(M).LT RMIN(1))GO TO 404
      GO TO 411
  404 KR(1)=KR(1)+1
      RMIN(1)=RI(M)
      ZMINS(1)=Z(M)
      GO TO 411
  405 IF(RI(M).LT.RMIN(2))GO TO 406
      GO TO 411
  406 KR(2)=KR(2)+1
      RMIN(2)=RI(M)
      ZMINS(2)=Z(M)
      GO TO 411
  407 IF(RI(M).LT.RMIN(3))GO TO 408
      GO TO 411
  408 KR(3)=KR(3)+1
      RMIN(3)=RI(M)
      ZMINS(3)=Z(M)
      GO TO 411
  409 IF(RI(M).LT.RMIN(4))GO TO 410
      GO TO 411
  410 KR(4)=KR(4)+1
      RMIN(4)=RI(M)
      ZMINS(4)=Z(M)
  411 CONTINUE
      S1=0.
      S2=0.
      KS=0
      DO 412 M=1, 4
      IF(KR(M) EQ 0)GO TO 412
      S1=S1+ZMINS(M)/(RMIN(M)*RMIN(M))
      S2=S2+1/(RMIN(M)*RMIN(M))
      KS=KS+1
  412 CONTINUE
      IF(KS.EQ.0)GO TO 413
      ZZ=S1/S2
      GO TO 415
```

```
413 ZZ=-10E10
    GO TO 415
414 ZZ=Z(M)
415 RETURN
    END
```

(9)绝对坐标变换为相对坐标的子程序

```
      SUBROUTINE MP1(N, MP, X0, Y0)
      COMMON X(500),Y(500),Z(500),H(11),ZN(500),ZQ(500),AZ(66,67),
     -III(66), JJJ(66)
      PM0=100.0/FLOAT(MP)
      DO 1 I=1,N
      X(I)=(X(I)-X0)*PM0
    1 Y(I)=(Y(I)-Y0)*PM0
      END
```

(10)相对坐标变换为绝对坐标的子程序

```
      SUBROUTINE MP2(N, MP, X0, Y0)
      COMMON X(500), Y(500), Z(500), H(11), ZN(500), ZQ(500),
     -AZ(66, 67), III(66), JJJ(66)
      PM0=100.0/FLOAT(MP)
      DO 1 I=1, N
      X(I)=X(I)/PM0+X0
    1 Y(I)=Y(I)/PM0+Y0
      END
```

(11)打印上图框子程序

```
      SUBROUTINE WR1(MM, MN, L)
      WRITE(2, 100)MM, MN, L
  100 FORMAT(15X,'比例尺=1/',I7,2X,'[图号',I1,'——',I1,']'/10X, 114
     -(1H-)/10X, 1HI, 56(2H( )), 1HI/10X, 3HI( ), 108(1H-), 3H( )I/10X,
      4HI( )I, 106(1H), 4HI( )I/10X, 6HI( )I, 102(1H-), 6H I( )I)
      END
```

(12)打印下图框子程序

```
      SUBROUTINE WR2
      WRITE(2, 100)
  100 FORMAT(10X, 6HI( )I , 102(1H-), 6H I( )I/10X, 4HI( )I,106(1H),
     -4HI( )I/10X,3HI( ),108(1H-),3H( )I/10X, 1HI, 56(2H( )), 1HI/10X, 114
      (1H-))
      END
```

六、多项式趋势分析计算结果

原图比例尺：100000

原图横向长度：6.40
原图纵向长度：6.40
接图次数：1
平面图分带数：10
平面插值计算方式：1
平面图的扫描行数：43
坐标选用方式：0
是否作残差趋势分析：0
趋势分析最低次数：1
趋势分析最高次数：1

趋势分析原始数据表

序号	横坐标	纵坐标	观测值
1	.300	6.100	870.000
2	1.400	6.200	793.000
3	2.400	6.100	755.000
4	3.600	6.200	690.000
5	5.700	6.200	800.000
6	1.600	5.200	800.000
7	2.900	5.100	730.000
8	3.400	5.300	728.000
9	3.400	5.700	710.000
10	4.800	5.600	780.000
11	5.300	5.000	804.000
12	6.200	5.200	855.000
13	.200	4.300	830.000
14	.900	4.200	813.000
15	2.300	4.800	762.000
16	2.500	4.500	765.000
17	3.000	4.500	740.000
18	3.500	4.500	765.000
19	4.100	4.600	760.000
20	4.900	4.200	790.000
21	6.300	4.300	820.000
22	.900	3.200	855.000
23	1.700	3.800	812.000
24	2.400	3.800	773.000
25	3.700	3.500	812.000
26	4.500	3.200	827.000
27	5.200	3.200	805.000

28	6.300	3.400	840.000
29	.300	2.400	890.000
30	2.000	2.700	820.000
31	3.800	2.300	873.000
32	6.300	2.200	875.000
33	.600	1.700	873.000
34	1.500	1.800	865.000
35	2.100	1.800	841.000
36	2.100	1.100	862.000
37	3.100	1.100	908.000
38	4.500	1.800	855.000
39	5.500	1.700	850.000
40	5.700	1.000	882.000
41	6.200	1.000	910.000
42	.400	.500	940.000
43	1.400	.600	915.000
44	1.400	.100	890.000
45	2.100	.700	880.000
46	2.300	.300	870.000
47	3.100	.010	880.000
48	4.100	.800	960.000
49	5.400	.400	890.000
50	6.000	.100	860.000
51	5.700	3.000	830.000
52	3.600	6.000	705.000

原始数据点位置图及观测值平面插值图从略，请参阅正文。

趋势方程X项系数方次：

0 1 0

趋势方程Y项系数方次：

0 0 1

观测值平均值：827.077

趋势方程系数矩阵AZ（1）：

52.000	172.600	167.010	43008.000
172.600	752.700	554.921	142434.100
167.010	554.921	737.500	133049.700

1次趋势方程有解：KIJ=0

趋势方程系数

```
* * * * * * * * * * * * * * *
*                           *
*B(0)=             913.824  *
*B(1)=             -1.695   *
*B(2)=             -25.257  *
*                           *
* * * * * * * * * * * * * * *
```

趋势分析计算结果

序号	横坐标	纵坐标	观测值	趋势值	残差值
1	.300	6.100	870.000	759.246	110.754
2	1.400	6.200	793.000	754.855	38.145
3	2.400	6.100	755.000	755.686	-.686
4	3.600	6.200	690.000	751.125	-61.125
5	5.700	6.200	800.000	747.565	52.435
6	1.600	5.200	800.000	779.773	20.227
7	2.900	5.100	730.000	780.095	-50.095
8	3.400	5.300	728.000	774.196	-46.196
9	3.400	5.700	710.000	764.093	-54.093
10	4.800	5.600	780.000	764.245	15.755
11	5.300	5.000	804.000	778.552	25.448
12	6.200	5.200	855.000	771.974	83.026
13	.200	4.300	830.000	804.879	25.121
14	.900	4.200	813.000	806.218	6.782
15	2.300	4.800	762.000	788.690	-26.690
16	2.500	4.500	765.000	795.928	-30.928
17	3.000	4.500	740.000	795.080	-55.080
18	3.500	4.500	765.000	794.232	-29.232
19	4.100	4.600	760.000	790.689	-30.689
20	4.900	4.200	790.000	799.436	-9.436
21	6.300	4.300	820.000	794.536	25.464
22	.900	3.200	855.000	831.475	23.525
23	1.700	3.800	812.000	814.964	-2.964
24	2.400	3.800	773.000	813.777	-40.777
25	3.700	3.500	812.000	819.150	-7.150
26	4.500	3.200	827.000	825.371	1.629
27	5.200	3.200	805.000	824.185	-19.185
28	6.300	3.400	840.000	817.268	22.732
29	.300	2.400	890.000	852.698	37.302
30	2.000	2.700	820.000	842.239	-22.239

31	3.800	2.300	873.000	849.290	23.710
32	6.300	2.200	875.000	847.577	27.423
33	.600	1.700	873.000	869.869	3.131
34	1.500	1.800	865.000	865.818	-.818
35	2.100	1.800	841.000	864.801	-23.801
36	2.100	1.100	862.000	882.481	-20.481
37	3.100	1.100	908.000	880.785	27.215
38	4.500	1.800	855.000	860.732	-5.732
39	5.500	1.700	850.000	861.562	-11.562
40	5.700	1.000	882.000	878.903	3.097
41	6.200	1.000	910.000	878.055	31.945
42	.400	.500	940.000	900.517	39.483
43	1.400	.600	915.000	896.296	18.704
44	1.400	.100	890.000	908.925	-18.925
45	2.100	.700	880.000	892.583	-12.583
46	2.300	.300	870.000	902.347	-32.347
47	3.100	.010	880.000	908.316	-28.316
48	4.100	.800	960.000	886.667	73.333
49	5.400	.400	890.000	894.566	-4.566
50	6.000	.100	860.000	901.126	-41.126
51	5.700	3.000	830.000	828.388	1.612
52	3.600	6.000	705.000	756.177	-51.177

总离差平方和：　196029.700

残差平方和：　67171.340

回归平方和：　128858.400

拟合度：　.657

F检验值：　47.000

趋势图及残差图从略，请参阅正文。

程序九　Q型聚类分析

一、程序主要功能

本程序是按聚合法对样品进行分类的Q型聚类分析。所谓聚合法是指开始时每个样品自成一类，然后按某种分类统计量使最亲近的类合并为样品集团，使类的数目减少，直到所有样品成为一类为止。分类结果可用分类谱系图（聚类树）表示样品之间的亲疏关系。

（1）本程序给出的分类统计量有三种，即：距离系数，相似系数，相关系数。用户可以任选其中的一种。

（2）对变量的标准化有三种方法，即：标准差标准化，极差标准化，极差正规化。用户可以任选其中的一种。

（3）类与类合并时的分类统计量是按加权平均法进行计算的。

（4）本程序的计算结果存入输出文件中，可以由宽行打印机输出，包括：原始数据表，

样品分类结果及样品的分类谱系图。

二、程序符号说明

N——————————样品数（N应小于等于500）；

M——————————变量数（M应小于等于50）；

N与M的上述含义仅限于主程序，而子程序中的含义对调；

X(500,50)——————原始数据矩阵，矩阵的行号为样品编号，列号为变量编号；

K1 —————————原始数据的标准化方式；

K1=1为标准差标准化；

K1=2为极差标准化；

K1=3为极差正规化；

K2 —————————分类统计量的选择方式：

K2=1为距离系数；

K2=2为相似系数；

K2=3为相关系数；

KK(500)———————类的权系数，即：每类（样品集团）中包含的样品数；

XA(500) ——————原始数据矩阵的行（即样品）平均值；

XC(500)———————原始数据矩阵的行（即样品）方差（K2=3），或行（即样品）平方和（K2=2）；

LI,LJ————————聚类过程中每次被合并的两个类号，这里规定样品集团的类号为集团中的最小样品号；

M1 —————————分类谱系图中的垂向连线数；

M2 —————————分类谱系图中水平连线的总数；

M3 —————————分类谱系图的打印扫描行数；

X1(1000) ——————谱系图中水平连线的左端点横坐标；

X2(1000) ——————谱系图中水平连线的右端点横坐标；

KX1(500) ——————谱系图中水平连线的行号；

KX2(500) ——————谱系图中水平连线右端点向下连线的延续行数；

KM(500) ——————谱系图中自上而下的样品排序行号；

KN(500)———————并类过程中，样品重新排序时的换行工作单元；

XLJ(500) ——————每次并类后，生成新类的水平连线左端点横坐标；

KLJ(500) ——————每次并类后，生成新类的行号；

XS(101) ———————谱系图中分类统计量的刻度尺；

W(101) ———————打印谱系图的扫描工作单元；

KW(101) ——————谱系图中垂直连线延续行数。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D3-1.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
N, M, K1, K2
((X(I, J), J=1, M), I=1, N)
```

例如，下面的数据文件（D3-1.DAT）就是一个供用户检测本程序的数据文件：

```
11, 5, 3, 3
1., 1., 16., 1000., 2.4,
2., 3., 10.8, 2000., 3.53,
2., 4., 7., 500., 9.3,
3., 7., 20., 26., 1.4,
3., 7., 22., 87., 1.9,
6., 3., 28.1, 44.7, 3.7,
1., 3., 11., 3., 11.2,
2., 1., 10.1, 600., 25.8,
3., 1., 7., 48.4, 1.23,
1., 3., 8., 102., 1.39,
3., 1., 16., 2.5, 2.5
```

四、计算结果输出

本程序输出文件的约定名为D3-1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、Q型聚类分析主程序

```
      PROGRAM C9006
      DIMENSION X(500, 50)
      CHARACTER FILENA*20, NOYES
    1 WRITE(*,'(1X, 28(1H  )\)')
      WRITE(*,'(1X,A)')'Q型聚类分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名D3-1.DAT]: '
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='D3-1.DAT'
      OPEN(1, FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名[约定名D3-1.WRI]: '
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='D3-1.WRI'
      OPEN(2, FILE=FILENA, STATUS='NEW')
      WRITE(*,*)'开始读入聚类分析的原始数据；'
      READ(1,*,ERR=3)N, M, K1, K2
      READ(1,*ERR=3)((X(I,J),J=1,M),I=1,N)
```

```
      WRITE(*,*)'正在进行计算，请等待！'
      WRITE(2,*)'                 ***************'
      WRITE(2,*)'                 *             *'
      WRITE(2,*)'                 *  Q型聚类分析计算结果  *'
      WRITE(2,*)'                 *             *'
      WRITE(2,*)'                 ***************'
      WRITE(2, 101)N, M, K1, K2
  101 FORMAT(//5X,'样品数：',I3/5X,'变量数：',I3/5X,'标准化方式：', I1/
     -5X,'分类统计量：', I1)
      WRITE(2,'(//10X,A)')'原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,102)'样品序号', ('变量', J, J=1,MXY)
  102 FORMAT(/5X,A, 10(4X, A, I2, 1X))
      DO 2 I=1,N
    2 WRITE(2,103)I, (X(I,J), J=1, M)
  103 FORMAT(9X, I3, 1X, 10(F10.3, 1X)/13X, 10(F10.3, 1X)))
      CALL QXJL(N, M, K1, K2, X)
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    3 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
    Q型聚类分析子程序
      SUBROUTINE QXJL(M, N, K1, K2, X)
      DIMENSION X(500,50), KK(500), XA(500), XC(500), X1(1000),
     -X2(1000), KX1(1000), KX2(1000), KM(500), KN(500), XLJ(500),
     -KLJ(500), XS(101), KW(102)
      CHARACTER A, B, C, W(102)
      DATA A, B, C/' ','-','I'/
C     根据K1的选值，对变量进行标准化变换；
      IF(K1-2)1, 6, 10
C     如果K1=1，则对变量进行"标准差标准化"变换；
    1 DO 5 J=1,N
      S1=0.
```

```
      DO 2 I=1,M
    2 S1=S1+X(I,J)
      S1=S1/M
      S2=0.
      DO 3 I=1,M
    3 S2=S2+(X(I,J)-S1)**2
      S2=SQRT(S2/M)
      DO 4 I=1,M
    4 X(I,J)=(X(I,J)-S1)/S2
    5 CONTINUE
      GO TO 15
C     如果K1=2，则对变量进行“极差标准化”变换；
    6 DO 9 J=1,N
      XMAX=X(1,J)
      XMIN=X(1,J)
      DO 7 I=2,M
      IF(X(I,J).GT.XMAX)XMAX=X(I,J)
      IF(X(I,J).LT.XMIN)XMIN=X(I,J)
    7 CONTINUE
      S1=XMAX-XMIN
      DO 8 I=1,M
    8 X(I,J)=(X(I,J)-XMIN)/S1
    9 CONTINUE
      GO TO 15
C     如果K1=3，则对变量进行“极差正规化”变换；
   10 DO 14 J=1,N
      S1=0.
      DO 11 I=1,M
   11 S1=S1+X(I,J)
      S1=S1/M
      XMAX=X(1,J)
      XMIN=X(1,J)
      DO 12 I=2,M
      IF(X(I,J).GT.XMAX)XMAX=X(I,J)
      IF(X(I,J).LT.XMIN)XMIN=X(I,J)
   12 CONTINUE
      S2=XMAX-XMIN
      DO 13 I=1,M
   13 X(I,J)=(X(I,J)-S1)/S2
```

```
   14 CONTINUE
C     使如下工作单元处于初始化状态;
   15 DO 16 I=1,M
      KK(I)=1
      KM(I)=0
      KN(I)=0
      XA(I)=0.
      XC(I)=0.
      KLJ(I)=0
   16 XLJ(I)=0.
C     根据K2的选值，为计算分类统计量作准备;
      IF(K2-2)17,18,20
C     如果K2=1,令: YS=-10.E10
   17 YS=-10.E10
      GO TO 23
C     如果K2=2则计算数据矩阵的行(样品)平方和，为计算相似系数作准备，
C     并且令: YS=10.E10;
   18 YS=10.E10
      DO 19 I=1,M
      DO 19 J=1,N
   19 XC(I)=XC(I)+X(I,J)**2
      GO TO 23
C     如果K2=3，则计算数据矩阵的行(样品)平均值及离差平方合，为计算相关系
C     数作准备，并且令: YS=10.E10;
   20 YS=10.E10
      DO 22 I=1,M
      DO 21 J=1,N
   21 XA(I)=XA(I)+X(I,J)
      XA(I)=XA(I)/N
      DO 22 J=1,N
   22 XC(I)=XC(I)+(X(I,J)-XA(I))**2
C     M1为谱系图中的垂向连线数，M2为谱系图中的水平连线数，M3为谱系图的打印
C     扫描行数;
   23 M1=M-1
      M2=M*2-2
      M3=M*2-1
C     下面开始进行样品分类计算;
      WRITE(2,'(///14X,A/)')'Q型聚类分析计算结果'
      DO 78 L=1,M1
```

```
C     对工作单元赋初值;
      L1=0
      L2=0
      L3=0
      L4=0
      L5=(L-1)*2
      N2=0
      N1=0
C     对YM赋初值，为并类作准备;
      IF(K2.EQ.1)THEN
            YM=10.E10
      ELSE
            YM=-10.E10
      ENDIF
C     开始计算分类统计量，只需计算下三角部分;
      DO 31 I=2,M
      IF(KK(I).EQ.0) GO TO 31
      I1=I-1
      DO 30 J=1,I1
      IF(KK(J).EQ.0) GO TO 30
      S1=0.
C     根据K2的选值，计算分类统计量;
      IF(K2-2)24, 26, 28
C     如果K2=1则以"距离系数"为分类统计量;
   24 DO 25 K=1,N
   25 S1=S1+(X(J,K)-X(I,K))**2
      S=SQRT(S1)
      IF(S.LT.YM)THEN
          YM=S
          LI=I
          LJ=J
      ENDIF
      GO TO 30
C     如果K2=2则以"相似系数"为分类统计量;
   26 DO 27 K=1,N
   27 S1=S1+X(J,K)*X(I,K)
      S2=SQRT(XC(J)*XC(I))
      S=S1/S2
      IF(S.GT.YM)THEN
```

```
          YM=S
          LI=I
          LJ=J
       ENDIF
       GO TO 30
C      如果K2=3则以“相关系数”为分类统计量;
    28 DO 29 K=1,N
    29 S1=S1+(X(J, K)-XA(J))*(X(I, K)-XA(I))
       S2=SQRT(XC(J)*XC(I))
       S=S1/S2
       IF(S.GT.YM)THEN
          YM=S
          LI=I
          LJ=J
       ENDIF
    30 CONTINUE
    31 CONTINUE
C      YM1为第一次(L=1)并类时的分类统计量，YM2为最后一次(L=M1)并类
C      时的分类统计量，为后面计算谱系图的刻度尺作准备;
       IF(L.EQ.1)YM1=YM
       IF(L.EQ.M1)YM2=YM
C      输出：每次并类的两个类号及分类统计量;
       IF(K2.EQ.1)THEN
          IF(YM.LT.YS)YM=YS
          WRITE(2, 101)LI, LJ, YM
       ELSE
          IF(YM.GT.YS)YM=YS
          IF(K2.EQ.2)WRITE(2,102)LI, LJ, YM
          IF(K2.EQ.3)WRITE(2,103)LI,LJ,YM
       ENDIF
   101 FORMAT(5X,'合并类(', I3,',',I3,')的距离系数=', F12.4)
   102 FORMAT(5X,'合并类(', I3,',',I3,')的相似系数=', F12.4)
   103 FORMAT(5X,'合并类(', I3,',',I3,')的相关系数=', F12.4)
C      按加权平均法计算合并后的新类(样品集团)的各个变量值;
       DO 32 J=1, N
    32 X(LJ,J)=(X(LJ,J)*KK(LJ)+X(LI,J)*KK(LI))/(KK(LI)+KK(LI))
C      根据K2的选值，为计算“新类”与“原有类”之间的分类统计量作准备;
       IF(K2-2)38, 33, 35
C      如果K2=1，直接转向标号37;
```

```
C     如果K2=2则计算数据矩阵的行（样品）平方和，为计算相似系数作准备；
   33 XC(LJ)=0.
      DO 34 J=1,N
   34 XC(LJ)=XC(LJ)+X(LJ,J)**2
      GO TO 38
C     如果K2=3，则计算数据矩阵的行（样品）平均值及离差平方合，为计算相关系
C     数作准备；
   35 XA(LJ)=0.
      DO 36 J=1, N
   36 XA(LJ)=XA(LJ)+X(LJ,J)
      XA(LJ)=XA(LJ)/N
      XC(LJ)=0
      DO 37 J=1,N
   37 XC(LJ)=XC(LJ)+(X(LJ,J)-XA(LJ))**2
C     完成一次并类之后，计算谱系图中各种连线的位置坐标，根据LI，LJ两类之间
C     的关系，可以按以下七种情况分别计算处理；
   38 DO 39 K=1, M
      IF(KM(K).EQ.LI)GO TO 41
      IF(KM(K).EQ.LJ)GO TO 60
      IF(KM(K).EQ.0 )GO TO 40
   39 CONTINUE
C     当LI，LJ两类均未出现时，可按如下处理；
   40 KM(K)=LI
      KM(K+1)=LJ
      KX1(L*2-1)=(K-1)*2+1
      X1(L*2-1)=YM1
      KX1(L*2)=(K-1)*2+3
      X1(L*2)=YM1
      GO TO 77
C     当LI类已出现，而LJ类未出现时，可按如下处理；
   41 DO 42 J=K,M
      IF(KM(J).EQ.LJ)GO TO 49
      IF(KM(J).EQ.0 )GO TO 43
   42 CONTINUE
   43 L1=K+1
      L2=J-1
      IF(L2.LT.L1)GO TO 48
      DO 44 J=L1,L2
   44 KN(J)=KM(J)
```

```
      DO 45 J=L1,L2
   45 KM(J+1)=KN(J)
      DO 46 J=1,L5
      IF(KX1(J).GE.L1*2-1)KX1(J)=KX1(J)+2
   46 CONTINUE
      DO 47 J=1,M
      IF(KLJ(J).GE.L1*2-1)KLJ(J)=KLJ(J)+2
   47 CONTINUE
C     当LI类已出现，而LJ类未出现，但是LI类后面没有其它类时，可按如下处理；
   48 KM(L1)=LJ
      KX1(L*2-1)=KLJ(LI)
      X1(L*2-1)=XLJ(LI)
      KX1(L*2)=L1*2-1
      X1(L*2)=YM1
      GO TO 77
C     当LI类先出现，LJ类后出现，并且LI与LJ两类相邻时，可按如下处理；
   49 IF(J-K.NE.KK(LJ))GO TO 50
      KX1(L*2-1)=KLJ(LI)
      X1(L*2-1)=XLJ(LI)
      KX1(L*2)=KLJ(LJ)
      X1(L*2)=XLJ(LJ)
      GO TO 77
C     当LI类先出现，LJ类后出现，而LI与LJ两类不相邻时，可按如下处理；
   50 L1=K+1
      L2=J-KK(LJ)
      L3=L2+1
      L4=J
      DO 51 J=L1,L4
   51 KN(J)=KM(J)
      N1=KK(LJ)
      DO 52 J=L1,L2
   52 KM(J+N1)=KN(J)
      N2=L4-K-N1
      DO 53 J=L3,L4
   53 KM(J-N2)=KN(J)
      DO 56 J=1,L5
      IF(KX1(J).GE.L1*2-1.AND.KX1(J).LE.L2*2-1)GO TO 54
      IF(KX1(J).GE.L3*2-1.AND.KX1(J).LE.L4*2-1)GO TO 55
      GO TO 56
```

```
54 KX1(J)=KX1(J)+N1*2
   GO TO 56
55 KX1(J)=KX1(J)-N2*2
56 CONTINUE
   DO 59 J=1,M
   IF(KLJ(J).GE.L1*2-1.AND.KLJ(J).LE.L2*2-1)GO TO 57
   IF(KLJ(J).GE.L3*2-1.AND.KLJ(J).LE.L4*2-1)GO TO 58
   GO TO 59
57 KLJ(J)=KLJ(J)+N1*2
   GO TO 59
58 KLJ(J)=KLJ(J)-N2*2
59 CONTINUE
   KX1(L*2-1)=KLJ(LI)
   X1(L*2-1)=XLJ(LI)
   KX1(L*2)=KLJ(LJ)
   X1(L*2)=XLJ(LJ)
   GO TO 77
C     当LJ类已出现，而LI类未出现时，可按如下处理；
60 DO 61 J=K,M
   IF(KM(J).EQ.LI)GO TO 67
   IF(KM(J).EQ.0 )GO TO 62
61 CONTINUE
62 L1=K-KK(LJ)+1
   L2=J-1
   DO 63 J=L1,L2
63 KN(J)=KM(J)
   DO 64 J=L1,L2
64 KM(J+1)=KN(J)
   KM(L1)=LI
   DO 65 J=1,L5
   IF(KX1(J).GE.L1*2-1) KX1(J)=KX1(J)+2
65 CONTINUE
   DO 66 J=1,M
   IF(KLJ(J).GE.L1*2-1)KLJ(J)=KLJ(J)+2
66 CONTINUE
   KX1(L*2-1)=L1*2-1
   X1(L*2-1)=YM1
   KX1(L*2)=KLJ(LJ)
   X1(L*2)=XLJ(LJ)
```

```
      GO TO 77
C     当LJ类先出现，LI类后出现时，可按如下处理；
   67 L1=K-KK(LJ)+1
      L2=J-KK(LI)
      L3=L2+1
      L4=J
      DO 68 J=L1,L4
   68 KN(J)=KM(J)
      N1=KK(LI)
      DO 69 J=L1,L2
   69 KM(J+N1)=KN(J)
      N2=L3-L1
      DO 70 J=L3,L4
   70 KM(J-N2)=KN(J)
      DO 73 J=1,L5
      IF(KX1(J).GE.L1*2-1.AND.KX1(J).LE.L2*2-1)GO TO 71
      IF(KX1(J).GE.L3*2-1.AND.KX1(J).LE.L4*2-1)GO TO 72
      GO TO 73
   71 KX1(J)=KX1(J)+N1*2
      GO TO 73
   72 KX1(J)=KX1(J)-N2*2
   73 CONTINUE
      DO 76 J=1,M
      IF(KLJ(J).GE.L1*2-1.AND.KLJ(J).LE.L2*2-1)GO TO 74
      IF(KLJ(J).GE.L3*2-1.AND.KLJ(J).LE.L4*2-1)GO TO 75
      GO TO 76
   74 KLJ(J)=KLJ(J)+N1*2
      GO TO 76
   75 KLJ(J)=KLJ(J)-N2*2
   76 CONTINUE
      KX1(L*2-1)=KLJ(LI)
      X1(L*2-1)=XLJ(LI)
      KX1(L*2)=KLJ(LJ)
      X1(L*2)=XLJ(LJ)
C     根据上面的计算结果，继续计算谱系图中各种连线的位置坐标；
   77 KX2(L*2-1)=KX1(L*2)-KX1(L*2-1)-1
      KX2(L*2)=0
      X2(L*2-1)=YM
      X2(L*2)=YM
```

```
      KLJ(LJ)=(KX1(L*2)+KX1(L*2-1))/2
      XLJ(LJ)=YM
      KK(LJ)=KK(LJ)+KK(LI)
      KK(LI)=0
      YS=YM
   78 CONTINUE
C     下面开始打印样品的分类谱系图;
      WRITE(2,'(/////52X,A)')'样品聚类谱系图'
      WRITE(2,'(9X,A)')'样品号'
C     YMS为分类统计量的变化范围;
      YMS=YM1-YM2
      YMD=YMS/100.0
C     XS为分类统计量的刻度值;
      DO 79 I=1,102
      XS(I)=YM1-YMD*(I-1)
   79 KW(I)=0
      KA=0
C     下面为扫描打印谱系图;
      DO 87 I=1,M3
      DO 80 J=1,102
      W(J)=A
      IF(KW(J).GT.0)W(J)=C
   80 KW(J)=KW(J)-1
      DO 85 J=1,M2
      IF(KX1(J).NE.I)GO TO 85
      IF(K2.EQ.1)GO TO 82
      DO 81 K=1,101
      IF(XS(K).LE.X1(J).AND.XS(K).GE.X2(J).AND.W(K).NE.C)W(K)
     -=B
      IF(XS(K+1).LE.X2(J))GO TO 84
   81 CONTINUE
      GO TO 85
   82 DO 83 K=1,101
      IF(XS(K).GE.X1(J).AND.XS(K).LE.X2(J).AND.W(K).NE.C)W(K)
     -=B
      IF (XS(K+1).GE.X2(J))GO TO 84
   83 CONTINUE
   84 IF(KX2(J).EQ.0)GO TO 85
      KW(K)=KX2(J)
```

```
 85 CONTINUE
    IF(MOD(1,2).EQ.1)GO TO 86
    WRITE(2,104)(W(J),J=1,101)
    GO TO 87
 86 KA=KA+1
    W(1)=B
    WRITE(2,105)KM(KA),(W(J),J=1,101)
 87 CONTINUE
    IF(K2.EQ.1)WRITE(2,106)(XS(I),I=1,101,10)
    IF(K2.EQ.2)WRITE(2,107)(XS(I),I=1,101,10)
    IF(K2.EQ.3)WRITE(2,108)(XS(I),I=1,101,10)
104 FORMAT(20X,101A1)
105 FORMAT(10X,I5,5(1H-),101A1)
106 FORMAT(//20X,1H+,10(10H----------+),2X,'距离系数'/15
   -X, 11F10.4)
107 FORMAT(//20X,1H+,10(10H----------+),2X,'相似系数'/15
   -X, 11F10.4)
108 FORMAT(//20X,1H+,10(10H----------+),2X,'相关系数'/15
   -X, 11F10.4)
    END
```

六、Q型聚类分析计算结果

样品数：11

变量数： 5

标准化方式：3

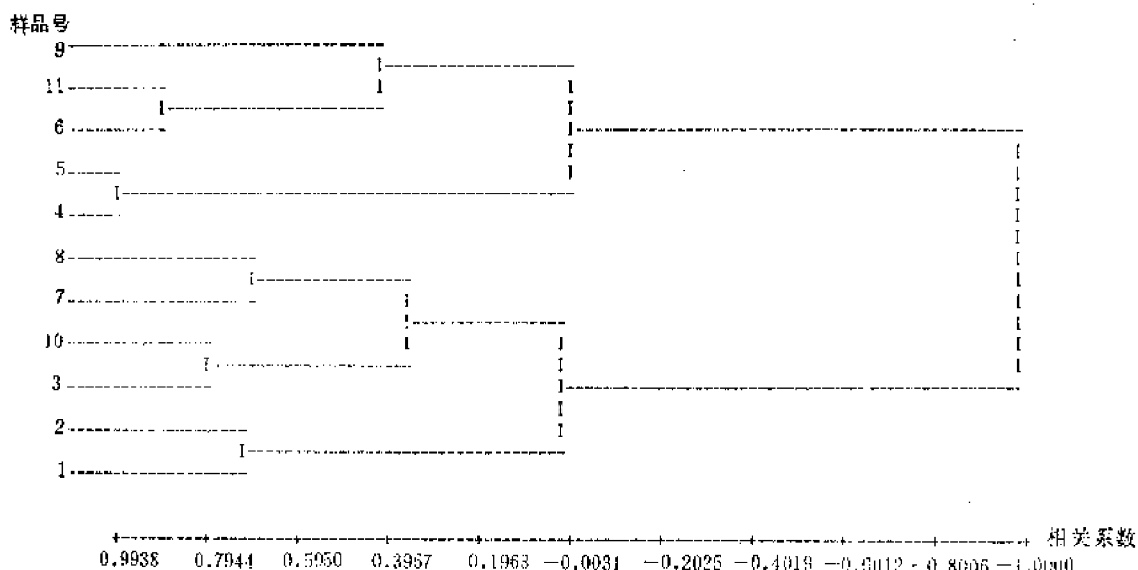
分类统计量：3

原始数据表

样品序号	变量1	变量2	变量3	变量4	变量5
1	1.000	1.000	16.000	1000.000	2.400
2	2.000	3.000	10.800	2000.000	3.530
3	2.000	4.000	7.000	500.000	9.300
4	3.000	7.000	20.000	26.000	1.400
5	3.000	7.000	22.000	87.000	1.900
6	6.000	3.000	28.100	44.700	3.700
7	1.000	3.000	11.000	3.000	11.200
8	2.000	1.000	10.100	600.000	25.800
9	3.000	1.000	7.000	48.400	1.230
10	1.000	3.000	8.000	102.000	1.390
11	3.000	1.000	16.000	2.500	2.500

Q型聚类分析计算结果

合并类（ 5，4 ）的相关系数＝ .9938
合并类（11，6 ）的相关系数＝ .8762
合并类（10，3 ）的相关系数＝ .7880
合并类（ 2，1 ）的相关系数＝ .7044
合并类（ 8，7 ）的相关系数＝ .6765
合并类（ 9，6 ）的相关系数＝ .4083
合并类（ 7，3 ）的相关系数＝ .3394
合并类（ 3，1 ）的相关系数＝ .0026
合并类（ 6，4 ）的相关系数＝ －.0146
合并类（ 4，1 ）的相关系数＝ －1.0000



附图7 样品聚类谱系图

程序十 两组判别分析

一、程序主要功能

两组判别分析所讨论的问题是指样品的归属类型只有两种，其原理是根据归属类型已确知的两组样品建立判别方程，即：

$$Y=C_1X_1+C_2X_2+\dots+C_mX_m$$

由判别方程可以确定两类样品的界线值。对于一个新样品，如果需要判别其类型归属，可将与归属类型有关的地质变量值代入判别方程，求得该样品的判别值；再根据这一判别值与类型归属的界线值进行比较，最后确定新样品的类型归属。

（1）要根据已知分类的样品（其中第一组的样品数为N1，第二组的样品数为N2）选择与归属类型有关的M个变量建立判别方程，建立方程后要确定样品类型归属的界线值YC。

（2）两组判别分析是假设样品取自两个不同母体，如果两个母体所选用的M个变量在统计上的差异不显著时，其判别结果显然是无意义的。

（3）本程序的计算结果存入输出文件中，可以由宽行打印机输出，包括：原始数据表，

判别方程的系数，类型归属的界线值，样品（类型归属）的判别图。

（4）本程序在计算时要进行F检验，必须在工作盘中有F.DAT数据文件。

二、程序符号说明

N1 ——————————已知属于第一组的样品数（要求N1小于等于250）；

N2 ——————————已知属于第二组的样品数（要求N2小于等于250）；

N3 ——————————未知分类的待判样品数（要求N3小于等于100）；

M ——————————变量数（要求M小于等于25）；

X1(250,25) —————第一组样品的原始数据；

X2(250, 25)—————第二组样品的原始数据；

X3(100, 25)—————待判样品的原始数据；

CP1(25)———————第一组样品每个变量的平均值；

CP2(25)———————第二组样品每个变量的平均值；

XS(25, 26) —————求解判别方程时的系数增广矩阵，矩阵的最后一列为判别方程的待定系数；

Y1(250)———————第一组样品的每个样品判别值；

Y2(250)———————第二组样品的每个样品判别值；

Y3(100)———————待判样品的每个样品判别值；

H(101) ———————样品判别图的刻度值。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D4-1.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
------------------------------------
N1, N2, N3, M
((X1(I, J), J=1, M), I=1, N1)
((X2(I,J),J=1,M),I=1,N2)
[((X3(I,J),J=1,M),I=1,N3)]
------------------------------------
```

例如，下面的数据文件（D4-1.DAT）就是一个供用户检测本程序的数据文件：

```
------------------------------------
2, 4, 1, 3
2.98, 0.31, 0.53, 3.2, 0.53, 0.77
2.53, 0.47, 0.49, 2.59, 0.30, 0.27, 2.96,
3.05, 1.50, 3.12, 2.84, 1.99
2.84, 0.60, 0.72
------------------------------------
```

四、计算结果输出

本程序输出文件的约定名为D4-1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、两组判别分析主程序

```
      PROGRAM C9009
      COMMON X1(250,25),X2(250,25),X3(100,25),CP1(25),CP2(25),
     -XS(25,26),Y1(250),Y2(250),Y3(100),H(101)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,28(1H  )\)')
      WRITE(*,'(1X,A)')'两组判别分析分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名D4-1.DAT]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='D4-1.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名[约定名D4-1.WRI]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='D4-1.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入两组判别分析的原始数据；'
      READ(1,*,ERR=2)N1,N2,N3,M
      READ(1,*,ERR=2)((X1(I,J),J=1,M),I=1,N1)
      READ(1,*,ERR=2)((X2(I,J),J=1,M),I=1,N2)
      IF(N3.NE.0)READ(1,*,ERR=2)((X3(I,J),J=1,M),I=1,N3)
      WRITE(*,*)'正在进行计算，请等待！'
      CALL LZPBFX(N1, N2, N3, M)
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    2 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

（1）两组判别分析子程序

```
      SUBROUTINE LZPBFX(N1,N2,N3,M)
      COMMON X1(250,25),X2(250,25),X3(100,25),CP1(25),CP2(25),
     -XS(25,26),Y1(250),Y2(250),Y3(100),H(101)
      CHARACTER W(100),A,B,C,D,E
      DATA A,B,C,D,E/'A','B',' ','|','I'/
      WRITE(2,*)'                 ***************'
      WRITE(2,*)'                 *             *'
```

```
      WRITE(2,*)'                *  两组判别分析计算结果  *'
      WRITE(2,*)'                *                        *'
      WRITE(2,*)'                ***************'
      WRITE(2,100)N1,N2,N3,M
  100 FORMAT(//5X,'第一组样品数：',I3/5X,'第二组样品数：',I3/5X,'待判
     -样品数：',I3/5X,'变量数：',I3)
C     计算已知第一，二组样品的变量均值CP1及CP2；
      DO 3 J=1,M
      S1=0.
      DO 1 I=1,N1
    1 S1=S1+X1(I,J)
      CP1(J)=S1/N1
      S1=0.
      DO 2 I=1,N2
    2 S1=S1+X2(I,J)
    3 CP2(J)=S1/N2
      WRITE(2,'(//10X,A)')'已知第一组样品原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      DO 4 I=1,N1
    4 WRITE(2,102)I,(X1(I,J),J=1,M)
  102 FORMAT(9X,I3,1X,10(F10.3,1X)/(13X,10(F10.3,1X)))
      WRITE(2,103)(CP1(J),J=1,M)
  103 FORMAT(5X,'变量均值',10(F10.3,1X)/(13X,10(F10.3,1X)))
      WRITE(2,'(//10X,A)')'已知第二组样品原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
      DO 5 I=1,N2
    5 WRITE(2,102)I,(X2(I,J),J=1,M)
      WRITE(2,103)(CP2(J),J=1,M)
      IF(N3.NE.0)THEN
        WRITE(2,'(//10X,A)')'待判样品原始数据表'
        MXY=M
        IF(M.GT.10)MXY=10
        WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
        DO 6 I=1,N3
```

```
   6   WRITE(2,102)I, (X3(I,J),J=1,M)
      ENDIF
C     计算判别方程的系数增广矩阵XS;
      M1=M+1
      DO 9 I=1,M
      DO 9 J=1,M
      S1=0.
      S2=0.
      DO 7 K=1,N1
   7  S1=S1+(X1(K,I)-CP1(I))*(X1(K,J)-CP1(J))
      DO 8 K=1,N2
   8  S2=S2+(X2(K,I)-CP1(I))*(X2(K,J)-CP2(J))
   9  XS(I,J)=S1+S2
      DO 10 I=1,M
  10  XS(I,M1)=CP1(I)-CP2(I)
C     调用高斯消元法子程序，求解判别方程的系数。EP为求解时的计算精度。求解
C     后，矩阵的最后一列为判别方程的系数;
      EP=10E-10
      CALL GAUSS(M,M1,EP,KIJ)
      IF(KIJ.EQ.0)THEN
        WRITE(2,104)KIJ
      ELSE
        WRITE(2,105)KIJ
        RETURN
      ENDIF
 104  FORMAT(//5X,判别方程有解! ', 5X,'KIJ=', I1)
 105  FORMAT(//5X,判别方程无解! ', 5X,'KIJ=', I1)
C     输出判别方程的系数;
      WRITE(2,106)
 106  FORMAT(//19X,'判别方程系数'/10X, 30(1H*)/10X, 1H*, 28(1H),
      1H*)
      DO 12 I=1,M
  12  WRITE(2,107)I,XS(I,M1)
 107  FORMAT(10X,4H*   C(,I2,2H)=,F20.5,2H   *)
      WRITE(2,108)
 108  FORMAT(10X,1H*,28(1H  ),1H*/10X,30(1H*))
C     计算第一，二组样品的综合指标YA，YB，以及样品类型归属的界线值YC;
      YA=0.
      YB=0.
```

```
      DO 13 I=1,M
      YA=YA+CP1(I)*XS(I,M1)
   13 YB=YB+CP2(I)*XS(I,M1)
      YC=(N1*YA+N2*YB)/(N1+N2)
      WRITE(2,109)YA,YB,YC
  109 FORMAT(//5X,'第一组样品的综合指标：', F10.3/5X,
     -'第二组样品的综合指标：', F10.3/5X, '样品类型归属的界线值：', F10.3)
C     对判别方程进行显著性检验，D2为Mahalanobis距离，F0是由D2为基础构成的
C     检验统计量，F0服从F(M, N1+N2-M-1)分布，如果F0>F，则判别方程
C     有意义；如果F0<F，则判别方程无意义。
      D2=0.
      DO 14 I=1,M
   14 D2=D2+(CP1(I)-CP2(I))*XS(I,M1)
      D2=FLOAT(N1+N2-2)*D2
      F0=D2*FLOAT(N1*N2*(N1+N2-M-1))/FLOAT((N1+N2)*
     -(N1+N2-2)*M)
      IFF=N1+N2-M-1
      F=FJY(M,IFF)
      IF(F0.GT.F)THEN
        WRITE(2,110)D2,F0,M,IFF,F
      ELSE
        WRITE(2,111)D2,F0,M,IFF,F
      ENDIF
  110 FORMAT(//5X,'马哈拉诺比斯距离：', F10.3/5X, '统计量F0=',F10.3,
     -'>F(',I2, ',', I2, ')=', F10.3)
  111 FORMAT(//5X,'马哈拉诺比斯距离：', F10.3/5X, '统计量F0=',F10.3,
     -'<F(', I2, ',', I2, ')=', F10.3)
C     根据判别方程，计算每个样品的判别值，Y1，Y2分别为第一，二组样品的判别
C     值，Y3为待判样品的判别值。
      DO 15 I=1,N1
      Y1(I)=0.
      DO 15 J=1,M
   15 Y1(I)=Y1(I)+X1(I,J)*XS(J,M1)
      DO 16 I=1,N2
      Y2(I)=0.
      DO 16 J=1,M
   16 Y2(I)=Y2(I)+X2(I,J)*XS(J,M1)
      IF(N3.NE.0)THEN
        DO 17 I=1,N3
```

```
      Y3(I)=0
      DO 17 J=1,M
17    Y3(I)=Y3(I)+X3(I,J)*XS(J,M1)
      ENDIF
C     确定全部样品判别值的最大值YMAX及最小值YMIN;
      YMAX=Y1(1)
      YMIN=Y1(1)
      DO 18 I=1,N1
      IF(Y1(I).GT.YMAX)YMAX=Y1(I)
      IF(Y1(I).LT.YMIN)YMIN=Y1(I)
18    CONTINUE
      DO 19 I=1,N2
      IF(Y2(I).GT.YMAX)YMAX=Y2(I)
      IF(Y2(I).LT.YMIN)YMIN=Y2(I)
19    CONTINUE
      IF(N3.NE.0)THEN
      DO 20 I=1,N3
         IF(Y3(I).GT.YMAX)YMAX=Y3(I)
         IF(Y3(I).LT.YMIN)YMIN=Y3(I)
20    CONTINUE
      ENDIF
C     计算样品判别图的刻度值H;
      DY=(YMAX-YMIN)/100
      YMAX=YMAX+10*DY
      YMIN=YMIN-10*DY
      DY=(YMAX-YMIN)/100
      DO 21 I=1,101
21    H(I)=YMIN+(I-1)*DY
C     确定类型归属界线值的刻度位置IYC;
      DO 22 I=1,100
      IF(YC.GE.H(I).AND.YC.LT.H(I+1))GO TO 23
22    CONTINUE
23    IYC=I
      WRITE(2,112)
112   FORMAT(/////5X,'  样品序号'4X,'判别函数值',45X,'样品判别图'//5X,'
     一第一组样品', 14X, '+', 10(10H----------+))
      DO 28 I=1,N1
      DO 27 J=1,100
      IF(Y1(I).GE.H(J).AND.Y1(I).LT.H(J+1))GO TO 24
```

```
      W(J)=C
      GO TO 26
   24 IF(Y1(I).GT.YC) GO TO 25
      W(J)=B
      GO TO 26
   25 W(J)=A
   26 IF(J.EQ.IYC.AND.W(J).EQ.C)W(J)=D
   27 CONTINUE
      W(100)=E
   28 WRITE(2,113)I,Y1(I),(W(J),J=1,100)
  113 FORMAT(12X,I3,3X,F10.3,'  I',100A1)
      WRITE(2,114)
  114 FORMAT(5X,'第二组样品', 14X, '+', 10(10H---------+))
      DO 33 I=1,N2
      DO 32 J=1,100
      IF(Y2(I).GE.H(J).AND.Y2(I).LT.H(J+1))GO TO 29
      W(J)=C
      GO TO 31
   29 IF(Y2(I).GT.YC)GO TO 30
      W(J)=B
      GO TO 31
   30 W(J)=A
   31 IF(J.EQ.IYC.AND.W(J).EQ.C)W(J)=D
   32 CONTINUE
      W(100)=E
   33 WRITE(2,113)I,Y2(I),(W(J),J=1,100)
      WRITE(2,115)
  115 FORMAT(5X,'    待判样品', 14X,'+', 10(10H---------+))
      IF(N3.NE.0)THEN
      DO 38 I=1,N3
      DO 37 J=1,100
      IF(Y3(I).GE.H(J).AND.Y3(I).LT.H(J+1))GO TO 34
      W(J)=C
      GO TO 36
   34 IF(Y3(I).GT.YC)GO TO 35
      W(J)=B
      GO TO 36
   35 W(J)=A
   36 IF(J.EQ.IYC.AND.W(J).EQ.C)W(J)=D
```

```
37 CONTINUE
   W(100)=E
28 WRITE(2,113)I,,Y3(I),(W(J),J=1,100)
   ENDIF
   WRITE(2,116)
116 FORMAT(29X,'+',10(10H----------+))
   WRITE(2,117)(H(I),I=1,101,10)
117 FORMAT(5X,'判别值刻度', 8X, 11F10.3)
   END
```

(2)高斯消元法子程序

```
   SUBROUTINE GAUSS(N,M,EP,KIJ)
   COMMON X1(250,25),X2(250,25),X3(100,25),CP1(25),CP2(25),
  -XS(25,26)Y1(250),Y2(250),Y3(100),H(101)
   N1=N-1
   DO 4 L=1,N1
   P=0.
   DO 1 I=L,N
   IF(ABS(XS(I,L)).LE.ABS(P))GO TO 1
   P=XS(I,L)
   I0=I
 1 CONTINUE
   IF(ABS(P).LE.EP)GO TO 7
   IF(I0.EQ.L)GO TO 3
   DO 2 J=L,M
   T=XS(L,J)
   XS(L,J)=XS(I0,J)
 2 XS(I0,J)=T
 3 P=1./P
   K=L+1
   DO 4 I=K,M
   XS(L,J)=XS(L,J)*P
   DO 4 I=K,N
 4 XS(I,J)=XS(I,J)-XS(I,L)*XS(L,J)
   XS(N,M)=XS(N,M)/XS(N,N)
   DO 6 L=1,N1
   I=N-L
   K=I+1
   P=0.
   DO 5  J=K,N
```

```
 5 P=P+XS(I,J)*XS(J,M)
 6 XS(I,M)=XS(I,M)-P
   KIJ=0
   GO TO 8
 7 KIJ=1
 8 CONTINUE
   END
```

（3）F检验子程序

```
   FUNCTION FJY(N1,N2)
   INTEGER NEWN1(120),NEWN2(120)
   REAL BIK(34,19)
   DATA NEWN1/1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 2*11, 3*12, 5*13,
  -4*14, 6*15, 10*16, 20*17, 60*18/
   DATA NEWN2/1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
  -16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 10*31,
  -20*32, 60*33/
   DATA MK/0/
   IF(MK.EQ.1)GO TO 1
   MK=1
   OPEN(5,FILE='F.DAT')
   READ(5,'(19F7.2)')((BIK(I,J),J=1,19),I=1,34)
   CLOSE(5)
 1 IN1=19
   IN2=34
   IF(N1.LE.120)IN1=NEWN1(N1)
   IF(N2.LE.120)IN2=NEWN2(N2)
   FJY=BIK(IN2, IN1)
   END
```

六、两组判别分析计算结果

第一组样品数：2

第二组样品数：4

待判样品数：1

变量数：3

已知第一组样品原始数据表

样品序号	变量1	变量2	变量3
1	2.980	.310	.530
2	3.200	.530	.770
变量均值	3.090	.420	.650

已知第二组样品原始数据表

样品序号	变量1	变量2	变量3
1	2.530	.470	.490
2	2.590	.300	.270
3	2.960	3.050	1.500
4	3.120	2.840	1.990
变量均值	2.800	1.665	1.063

待判样品原始数据表

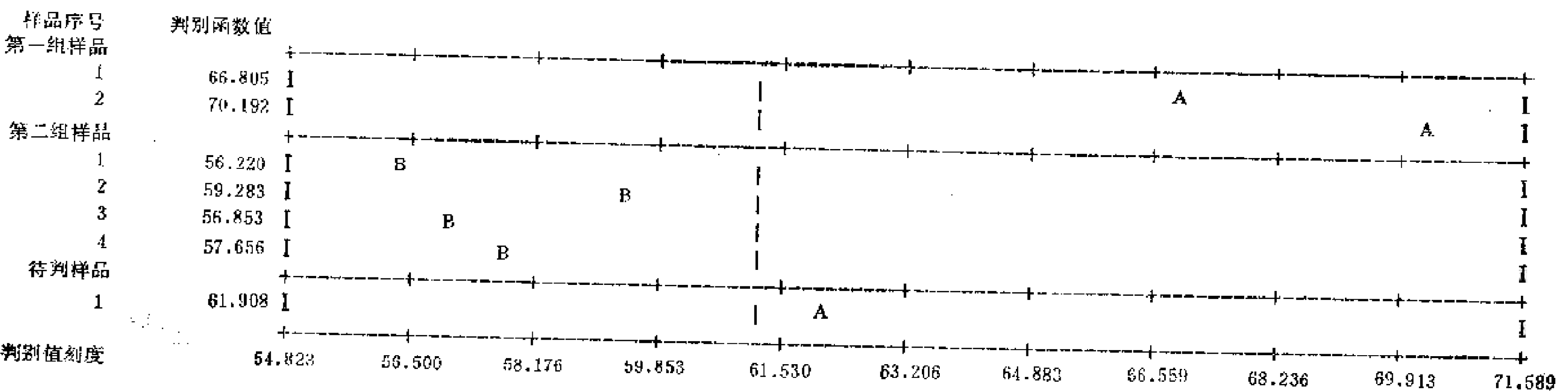
样品序号	变量1	变量2	变量3
1	2.840	.600	.720

判别方程有解：KIJ=0

判别方程系数

```
* * * * * * * * * * * * * * * *
*                             *
*  C(1)=         23.70643     *
*  C(2)=         -1.12773     *
*  C(3)=         -6.58558     *
*                             *
* * * * * * * * * * * * * * * *
```

第一组样品的综合指标：68.499
第二组样品的综合指标：57.503
样品类型归属的界线值：61.168
马哈拉诺比斯距离：43.982
统计量F0=9.774<F（3，2）=19.160



附图 8　两组判别分析的分类图

七、F检验时的数据文件

下面是两组判别分析进行F检验时的F.DAT数据文件，需要存入当前工作盘。

161.45	199.50	215.71	224.58	230.17	233.98	236.78	238.89	240.54	241.88	243.90	245.96	247.98	249.10	250.11	251.13	252.20	255.30	254.30
18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.96	5.66	5.63
6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.76	1.70	1.64
4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
4.08	3.23	2.84	2.61	2.45	2.34	2.35	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.06	1.61	1.55	1.50	1.43	1.35	1.25
3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

程序十一　逐步判别分析

一、程序主要功能

按着变量对判别问题的重要性，在计算过程中使变量有进有出，保留那些对划分样品类型归属起作用的变量，剔出那些对划分样品类型归属不起作用的变量或起作用不大的变量。

（1）逐步判别分析是假定各组的样品来自不同的母体，即：假定各母体之间相互独立，并且服从正态分布，各母体的数学期望不同，而协方差矩阵是相等的。

（2）为了筛选变量，需要计算每个变量的组（类）内离差矩阵W及总离差矩阵V，并以W/V作为引进或剔出变量的标准，如果已经计算了t步，在判别函数中引进了p个变量，则在t+1步要计算全部变量的判别效果，首先是考虑在已选入的变量中剔出判别效果最不显著的变量；如果不能剔出，则要从未引进的变量中找出判别效果最显著的变量进入判别函数，直到既不能剔出也不能引进变量时，逐步判别的计算过程结束。

（3）本程序的计算结果存入输出文件中，可以由宽行打印机输出，包括：原始数据表，引进或剔出变量的中间计算过程，各个已知样品组的分类结果以及待判样品的分类结果。

二、程序符号说明

M——————————变量数；

L——————————判别组数（要求L小于等于10），L+1组为待判样品组；

N——————————样品总数；

NI（11）—————— 各组样品数；

X（500，50）——— 原始数据矩阵，矩阵的行号为样品编号，列号为变量编号；

LJ(11,2)———————原始数据矩阵各组样品的行号，其中LJ(S，1)为第S组的起始行号，LJ（S，2）为终止行号；

GP（10，50）——— 各组样品的变量平均值；

GGP（50）——————全部已知分组样品的变量总平均值；

W（50，50）—————组（类）内离差矩阵；

V（50，50）———— 总的离差矩阵；

U（50）——————— 各变量的判别能力；

KL（50）—————— 变量是否引进的标志；

其中：引进的变量，令：KL=1，

未引进或剔除后的变量，令：KL=0；

LL——————————— 计算过程中，已引进的变量总数；

KJ——————————— 矩阵W及V的变换次数；

BI（10）——————— 分组符号；

EP——————————— 为防止除法运算时分母为零，要求分母的值大于EP，

此处令EP=10E−10；

QI（10）——————— 各组判别函数的先验概率；

COI（10）—————— 各组判别函数的COI项；

CAI(10,50)———— 各组判别函数的各个变量的CAI项；

FL（10）—————— 每个样品归属各组的概率；

LX (500) —————— 每个样品的归属组别；

MB (10, 10) ————— 各已知样品组判别后的分类矩阵；

KF ——————————— 用户给定F检验值时，令：KF=1，

自动选定F检验值时，令：KF=0；

AF1, AF2 —————— 分别为用户给定的引进与剔出变量时的F检验值；

FF ——————————— 程序中使用的F检验值。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D4-3.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
L,M,KF
(NI(I),I=1,L+1)
((X(I,J),J=1,M),I=1,N)
[AF1,AF2]
```

例如，下面的数据文件（D4-3.DAT）就是一个供用户检测本程序的数据文件：

```
3,4,1
7,4,6,0
6,-11.5,19,90,
-4,-15.0,13,54,
0,-23.0,5,-35,
-100,-21.4,7,-15,
-5,-18.5,15,18,
10,-18.0,14,50,
-8,-14.0,16,56,
90.2,-17.0,17,3,
0,-14.0,20,35,
-100,-21.5,15,-40,
13,-17.2,18,2,
-11,-18.5,25,-36,
0.5,-11.5,19,37,
-10,-19.0,21,-42,
20,-22.0,8,-20,
0.6,-13.0,26,21,
-40,-20.0,22,-50
2.0,2.0
```

四、计算结果输出

本程序输出文件的约定名为D4-3.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、逐步判别分析主程序

```
      PROGRAM C9011
      COMMON X(500,50),NI(11),LJ(11,2),CP(10,50),CCP(50),KL(50),
     -U(50),QI(10),COI(10),CAI(10,50),FL(10),LX(500),MB(10,10)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,28(1H  )\)')
       WRITE(*,'(1X,A)')'逐步判别分析分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名〔约定名D4-3.DAT〕：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='D4-3.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名〔约定名D4-3.WRI〕：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.'  ')FILENA='D4-3.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入逐步判别分析的原始数据：'
      READ(1,*,ERR=4)L,M,KF
      READ(1,*,ERR=4)(NI(I),I=1,L+1)
      IF(NI(L+1).EQ.0)THEN
        N=0
        DO 2 I=1,L
    2   N=N+NI(I)
        READ(1,*,ERR=4)((X(I,J),J=1,M),I=1,N)
        ELSE
        N=0
        DO 3 I=1,L+1
    3   N=N+NI(I)
        READ(1,*,ERR=4)((X(I,J),J=1,M),I=1,N)
      ENDIF
      WRITE(*,*)'正在进行计算，请等待！'
      WRITE(2,*)'              *****************'
      WRITE(2,*)'              *               *'
      WRITE(2,*)'              *  逐步判别分析计算结果  *'
      WRITE(2,*)'              *               *'
      WRITE(2,*)'              *****************'
      WRITE(2,100)L,M,N,KF
  100 FORMAT(//5X,'判别组数=',I3/5X,'变量个数=',I3/5X,'样品总
     -数=',I3/3X,'F检验方式=',I3)
      CALL ZBPBFX(L,M,KF)
```

```
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    4 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

(1) 逐步判别分析子程序

```
      SUBROUTINE ZBPBFX(L,M,KF)
      COMMON X(500,50),NI(11),LJ(11,2),CP(18,50),CCP(50),KL(50),
     -U(50),QI(10),COI(10),CAI(10,50),FL(10),LX(500),MB(10,10)
      DIMENSION W(50,50),V(50,50),WV(50,50),BI(10)
      CHARACTER A,B,C,RW,R,BI,WN,FM*100
C     给分组（类）符号赋值
      DATA BI/'1','2','3','4','5','6','7','8','9','0'/
      DATA A,B,C/'+','-',' '/
  100 FORMAT(//133('*'))
C     如果由用户给定F检验值，则要读入引进变量的F检验值AF1及剔出变量的F检
C     验值AF2.
      IF(KF.EQ.1)READ(1,*,ERR=48)AF1,AF2
C     给EP赋值，为防止除法运算时分母为零，要求分母的值大于EP.
      EP=10E-10
C     确定第一组样品在数据矩阵中的起始行号及终止行号.
      LJ(1,1)=1
      LJ(1,2)=NI(1)
C     确定第二组样品至待判组样品在数据矩阵中的起始行号及终止行号.
      IF(NI(L+1).EQ.0)THEN
        DO 1 I=2,L
        LJ(I,1)=LJ(I-1,1)+NI(I-1)
    1   LJ(I,2)=LJ(I-1,2)+NI(I)
      ELSE
        DO 2 I=1,L
        LJ(I+1,1)=LJ(I,1)+NI(I)
    2   LJ(I+1,2)=LJ(I,2)+NI(I+1)
      ENDIF
C     已知分组（类）样品的总数为N.
```

```
      N=LJ(L,2)
C     输出已知分组样品原始数据表与变量平均值；以及待判样品原始数据表、
      DO 6 K=1,L
      N1=LJ(K,1)
      N2=LJ(K,2)
C     计算各个已知分组样品的变量平均值CP.
      DO 4 J=1,M
      S1=0.
      DO 3 I=N1,N2
    3 S1=S1+X(I,J)
    4 CP(K,J)=S1/NI(K)
C     输出已知分组样品的原始数据表X.
      WRITE(2,101)K
  101 FORMAT(//10X,'第',I2,'组样品原始数据表')
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,102)'样品序号',('变量',J,J=1,MXY)
  102 FORMAT(/5X,A,10(4X,A,I2,1X))
      I1=0
      DO 5 I=N1,N2
      I1=I1+1
    5 WRITE(2,103)I1,(X(I,J),J=1,M)
  103 FORMAT(10X,I3,10(F10.3,1X)/(13X,10(F10.3,1X)))
C     输出已知分组样品的各变量平均值CP.
    6 WRITE(2,104)(CP(K,J),J=1,M)
  104 FORMAT(3X,'变量平均值',10(F10.3,1X))
      DO 8 J=1,M
      S1=0.
      DO 7 I=1,L
    7 S1=S1+CP(I,J)*NI(I)
    8 CCP(J)=S1/N
C     输出已知分组样品的全部变量总平均值CCP.
      WRITE(2,105)(CCP(I),I=1,M)
  105 FORMAT(/1X,'变量总平均值',10(F10.3,1X))
C     输出待判样品原始数据表.
      IF(NI(L+1).NE.0)THEN
        WRITE(2,'(//10X,A)')'待判样品原始数据表'
        MXY=M
        IF(M.GT.10)MXY=10
```

```
      WRITE(2,102)'样品序号',('变量',J,J=1,MXY)
      N1=LJ(L+1,1)
      N2=LJ(L+1,2)
      I1=0
      DO 9 I=N1,N2
      I1=I1+1
9     WRITE(2,103)I1,(X(I,J),J=1,M)
      ENDIF
C     计算组内离差矩阵W与总的离差矩阵V,因为W与V均为主对角线的对称矩
C     阵；所以，只需要计算矩阵W与V的上三角部分的元素值.
      DO 10 I=1,M
      DO 10 J=I,M
      W(I,J)=0.
      V(I,J)=0.
      DO 10 K=1,L
      N1=LJ(K,1)
      N2=LJ(K,2)
      DO 10 I1=N1,N2
      W(I,J)=W(I,J)+(X(I1,I)-CP(K,I))*(X(I1,J)-CP(K,J))
10    V(I,J)=V(I,J)+(X(I1,I)-CCP(I))*(X(I1,J)-CCP(J))
C     按对称关系求矩阵W与V下三角部分的元素值.
      DO 11 I=2,M
      I1=I-1
      DO 11 J=1,I1
      W(I,J)=W(J,I)
11    V(I,J)=V(J,I)
C     给KK,KJ,LL赋初值;其中KK=1时表示引进或剔出后，进行矩阵变换时出现分
C     母为零（即，调用子程序VARY时，W<EP,令KK=1）；KJ为矩阵W及V的
C     变换次数；LL为计算过程中已引进变量的总数.
      KK=0
      KJ=0
      LL=0
C     给各变量是否已引进的标志数组KL赋以初值零.
      DO 12 I=1,M
12    KL(I)=0
C     输出原始的组内离差矩阵W.
      WRITE(2,'(//10X,A)')'原始的组内离差矩阵'
      DO 13 I=1,M
13    WRITE(2,106)(W(I,J),J=1,M)
```

```
  106 FORMAT(5X,10F12.4)
C       输出原始的总离差矩阵V.
        WRITE(2,'(//10X,A)')'原始的总离差矩阵'
        DO 14 I=1,M
   14   WRITE(2,106)(V(I,J),J=1,M)
C       当：LL=M时，表示全部自变量都已引进判别函数中,并且转入标号3000,进行
C       判别方程的系数计算；否则进行引进变量的计算.
 1000   IF(LL.EQ.M)GO TO 3000
C       确定引进变量的F检验值FF1；当KF=0时，由函数子程序 FBIAO 计算产生
C       FF1值；当KF=1时，令FF=AF1，其中的AF1是由用户给定的引进变量的
C       F检验值.
        IF(KF.EQ.0)THEN
          IFF=N-L-LL
          JFF=L-1
          FF1=FBIAO(JFF,IFF)
        ELSE
          FF1=AF1
        ENDIF
C       输出引进变量的F检验值FF1.
        WRITE(2,107)FF1
  107   FORMAT(//10X,31('*')/10X,'*',29(' '),'*'/10X,'*',
     -2(' '),'引进变量的F检验值=',F6.2,2(' '),'*'/10X,'*',29(' '),
     -'*'/10X,31('*'))
C       为选择出判别能力最强的变量，给UMIN赋一个较大的初值.
        UMIN=10E10
C       分别计算已引进的变量与未引进的变量的判别能力.
        DO 17 I=1,M
C       如果KL(I)=1，说明第I个变量已引进判别函数中，则转入标号16进行计算.
        IF(KL(I).EQ.1)GO TO 16
C       如果出现V(I,I)小于EP则转入标号22，返回主程序，计算结束！
        IF(V(I,I).LT.EP)GO TO 22
C       计算并选择未引进变量中判别能力最强的变量，注意：U(I)的值越小判别能力
C       越强.
        U(I)=W(I,I)/V(I,I)
        IF(U(I).LT.UMIN)GO TO 15
        GO TO 17
   15   UMIN=U(I)
        KI=I
        GO TO 17
```

```
C       计算已引进变量的判别能力;
C       如果出现W(I,I)小于EP则转入标号23，返回主程序，计算结束!
   16   IF(W(I,I).LT.EP)GO TO 23
        U(I)=V(I,I)/W(I,I)
   17   CONTINUE
        FU=(1-UMIN)*(N-L-LL)/(UMIN*(L-1))
C       对未引进变量中判别能力最强的变量进行F检验，如果FU大于FF1时，则转入
C       标号18,引进第KI个变量；否则，输出检验结果后，转入标号3000，计算判
C       别函数值.
        IF(FU.GT.FF1)GO TO 18
        WRITE(2,'(//5X,A)')'此次计算结果表明，不能引进变量!'
        WRITE(2,108)KI,UMIN,FU,FF1
  108   FORMAT(//5X,即：未引进变量中第',I2,'号变量的判别能力最强,
     -U=',F10.4/5X,'检验统计量FU=',F10.4,'小于FF1=',F10.4,5X,
     -'因此，不能引进变量!')
        GO TO 3000
C       将第KI个变量引进判别函数，并且令KL(KI)=1;矩阵变换次数KJ加1；引进
C       变量的累计个数LL加1。
   18   KL(KI)=1
        KJ=KJ+1
        LL=LL+1
C       输出已引进的变量及未引进的变量的判别能力.
        WRITE(2,'(//5X,A)')'此次可以引进变量！计算结果如下：'
        WRITE(2,'(//5X,A)')'变量序号      变量判别能力      是否已引进'
        DO 19 I=1,M
        IF(KL(I).EQ.1)THEN
          WRITE(2,109)I,U(I),KL(I)
        ELSE
          WRITE(2,110)I,U(I),KL(I)
        ENDIF
   19   CONTINUE
  109   FORMAT(10X,I3,5X,F12.4,7X,'(',I1,')','已引进')
  110   FORMAT(10X,I3,5X,F12.4,7X,'(',I1,')','未引进')
        WRITE(2,111)KI,UMIN,FU,FF1,KI
  111   FORMAT(//5X,'即：未引进变量中第',I2,'号变量的判别能力最强,
     -U=',F10.4/5X,'检验统计量FU=',F10.4,'大于FF1=',F10.4,5X,'
     -因此引进第',I2,'号变量!')
        WRITE(2,100)
C       引进变量后，调用<矩阵元素变换子程序>进行矩阵W的元素变换.
```

```
      CALL VARY(W,WV,M,KI,EP,KK)
C     如果W矩阵变换时，出现分母为零则返回主程序，计算结束！
      IF(KK.EQ.1)RETURN
C     引进变量后，调用<矩阵元素变换子程序>进行矩阵V的元素变换.
      CALL VARY(V,WV,M,KI,EP,KK)
C     如果V矩阵变换时，出现分母为零则返回主程序，计算结束！
      IF(KK.EQ.1)RETURN
C     输出变换后的组内离差矩阵W
      WRITE(2,'(//10X,A,I3,A)')'第',KJ,'次变换后的组内离差矩阵'
      DO 20 I=1,M
   20 WRITE(2,106)(W(I,J),J=1,M)
C     输出变换后的总离差矩阵V.
      WRITE(2,'(//10X,A,I3,A)')'第',KJ,'次变换后的总离差矩阵'
      DO 21 I=1,M
   21 WRITE(2,106)(V(I,J),J=1,M)
C     如果引进的变量个数小于等于2，则继续引进变量；否则转入剔出变量计算.
      IF(KJ.LE.1)GO TO 1000
      GO TO 2000
C     矩阵V变换后，某个元素V(I,I)的值小于EP；返回主程序，计算结束！
   22 WRITE(2,112)KJ,I,I
  112 FORMAT(//5X,'第',I3,'次总的离差矩阵V变换后,元素V(',I2,',',I2,
     -')的值小于EP，返回主程序，计算结束！')
      RETURN
C     矩阵W变换后，某个元素W(I,I)的值小于EP；返回主程序，计算结束！
   23 WRITE(2,113)KJ,I,I
  113 FORMAT(//5X,'第',I3,'次总的离差矩阵W变换后，元素W(',I2,',',
     -I2,')的值小于EP，返回主程序，计算结束！')
      RETURN
C     如果LL=0时，表示所有的自变量都不能引进判别函数中；否则进行剔出变量
C     计算.
 2000 IF(LL.EQ.0)GO TO 3000
C     确定剔出变量的F检验值FF2；KF=0时，由函数子程序FBIAO计算产生
C     FF2值；KF=1时，令FF=AF2，其中的AF2是由用户给定的引进变量的
C     F检验值；
      IF(KF.EQ.0)THEN
        IFF=N-L-LL+1
        JFF=L-1
        FF2=FBIAO(JFF,IFF)
      ELSE
```

```
      FF2=AF2
      ENDIF
C     输出剔出变量的F检验值FF2;
      WRITE(2,114)FF2
  114 FORMAT(//10X,31('*')/10X,'*',29(' '),'*'/10X,'*',2(' '),
     -'剔出变量的F检验值=',F6.2,2(' ')'*'/10X,'*',29(' '),'*'/
     -10X,31('*'))
C     为选择出判别能力最弱的变量，给UMAX赋一个较大小的初值
      UMAX=-10E10
C     分别计算未引进的变量与已引进的变量的判别能力.
      DO 27 I=1, M
C     如果KL(I)=0，说明第I个变量未引进判别函数中，则转入标号26进行计算。
      IF(KL(I).EQ.0)GO TO 26
C     如果出现W(I, I)小于EP则转入标号32，返回主程序，计算结束!
      IF(W(I, I).LT.EP)GO TO 32
C     计算并选择已引进变量中判别能力最弱的变量，注意：U(I)的值越大判别能力越
C     弱.
      U(I)=V(I, I)/W(I, I)
      IF(U(I).GT.UMAX)GO TO 25
      GO TO 27
   25 UMAX=U(I)
      KI=I
      GO TO 27
C     计算未引进变量的判别能力,如果出现V(I,I)小于EP则转入标号33,返回主程序.
   26 IF(V(I, I).LT.EP)GO TO 33
      U(I)=W(I, I)/V(I, I)
   27 CONTINUE
C     对已引进变量中判别能力最弱的变量进行F检验。
      FU=(1-UMAX)*(N-L-LL+1)/(UMAX*(L-1))
C     如果FU小于FF2时，则转入标号28，剔出第KI个变量；否则，输出检验结果后，
C     转入标号1000，再进行引进变量计算。
      IF(FU.LT.FF2)GO TO 28
      WRITE(2, '(//5X, A)')'此次计算结果表明，不能剔出变量!'
      WRITE(2, 115)KI, UMAX, FU, FF2
  115 FORMAT(//5X, '即：已引进变量中第', I2, '号变量的判别能力最弱,
     -U=', F10.4/5X, '检验统计量FU=', F10.4, '大于FF2=', F10.4,
     -5X, '因此，不能剔出变量!')
      GO TO 1000
C     将第KI个变量从判别函数中剔出，并且令KL(KI)=0；矩阵变换次数KJ加1，引
```

```
C       进变量的累计个数LL减1。
     28 FL(KI)=0
        KJ=KJ+1
        LL=LL-1
C       输出已引进变量及未引进变量的判别能力。
        WRITE(2, '(//5X, A)')'此次可以剔出变量！计算结果如下：，
        WRITE(2, '(//5X, A)')'变量序号      变量判别能力      是否已引进'
        DO 29 I=1, M
          IF(KL(I).EQ.1)THEN
            WRITE(2, 109)I, U(I),KL(I)
          ELSE
            WRITE(2, 110)I, U(I), KL(I)
          ENDIF
     29 CONTINUE
        WRITE(2, 116)KI, UMIN, FU, FF2, KI
    116 FORMAT(//5X, '即：已引进 变 量 中 第', I2, '号变 量 的 判 别 能 力 最
       -弱，U=', F10.4/5X, '检 验 统 计 量FU=', F10.4, '小 于FF2=',
       -F10.4, 5X, '因此剔出第', I2, '号变量！')
        WRITE(2, 100)
C       剔出变量后，调用《矩阵元素变换子程序》进行矩阵W的元素变换。
        CALL VARY(W, WV, M, KI, EP, KK)
C       如果W矩阵变换时，出现分母为零则返回主程序，计算结束！
        IF(KK.EQ.1)RETURN
C       引进变量后，调用《矩阵元素变换子程序》进行矩阵V的元素变换。
        CALL VARY(V, WV, M, KI, EP, KK)
C       如果V矩阵变换时，出现分母为零则返回主程序，计算结束！
        IF(KK.EQ.1)RETURN
C       输出变换后的组内离差矩阵W。
        WRITE(2, '(//10X, A, I3, A)')'第',KJ,'次变换后的组内离差矩阵'
        DO 30 I=1, M
     30 WRITE(2, 106)(W(I, J), J=1, M)
C       输出变换后的总的离差矩阵V。
        WRITE(2, '(//10X, A, I3, A)')'第',KJ,'次变换后的总的离差矩阵'
        DO 31 I=1, M
     31 WRITE(2, 106)(V(I, J), J=1, M)
        GO TO 2000
C       出现总离差矩阵W变换后，元素W(I,I)的值小于EP；返回主程序，计算结束！
     32 WRITE(2, 113)KJ, I, I
        RETURN
```

```
C       出现总离差矩阵V变换后，元素V(I, I)的值小于EP；返回主程序，计算结束！
   33 WRITE(2, 112)KJ, I, I
      RETURN
C       当：LL=0时，表示所有自变量都不能引进方程中，转入标号4000，返回主程
C       序，计算结束！否则计算判别函数.
 3000 IF(LL.EQ.0)GO TO 4000
C       计算并输出各组的判别函数。
      WRITE(2, '(//10X, A)'), 各组判别函数值'
      DO 36 K=1, L
C       计算各组判别函数的先验概率QI项.
      QI(K)=NI(K)/FLOAT(N)
      S2=0.
      DO 35 I=1, M
      S1=0.
      IF(KL(I).EQ.0)GO TO 35
      DO 34 J=1, M
      IF(KL(J).EQ.0)GO TO 34
      S1=S1+W(I, J)*CP(K, J)
   34 CONTINUE
C       计算各组判别函数的CAI项。
      CAI(K, I)=S1*(N-L)
      S2=S2+S1*CP(K, I)
   35 CONTINUE
C       计算各组判别函数的COI项。
      COI(K)=S2*(N-L)*(-0.5)
   36 WRITE(2, 117)K, QI(K), COI(K), (I, CAI(K, I), I=1, M)
  117 FORMAT(/5X, '分组序号', 2X, '先验概率', 2X, 'COI项值', 2X, '变
     -量序号', 2X, 'CAI项值'/11X, I2, 2F10.3, 8X, I2, F10.3/(41X,
     -I2, F10.3))
C       输出已知分组样品及待判样品的分类结果；
      WRITE(2, '(//10X, A)')'已知分组样品及待判样品的分类结果如下：'
      DO 42 K=1, L+1
      IF(K.EQ.L+1.AND.NI(L+1).EQ.0)GQ TO 42
      IF(K.NE.L+1)THEN
        WRITE(2, 118)K, (KI, KI=1, L)
  118 FORMAT(/10X, '第', I2, '组样品的分类表'/2X, '样品序号', 2X, '
     -原始分组', 2X, '判别分组', 2X, '分组正误', 2X, '最大f值', 2X, '所
     -属分组', 7(2X, I2, '组f值'))
      ELSE
```

```
      WRITE(2, 119)(KI, KI=1, L)
  119 FORMAT(/10X, '待判样品的分类表'/2X,'样品序号', 2X, '原始分组',
     -2X, '判别分组', 2X, '分组正误', 2X, '最大f 值', 2X, '所属分组',
     -7(2X, I2, '组f值'))
      ENDIF
      N1=LJ(K, 1)
      N2=LJ(K, 2)
      KI=0
      R=BI(K)
      DO 41 I1=N1, N2
      KI=KI+1
      RW=B
      DO 38 I=1, L
      S1=0.
      DO 37 J=1, M
      IF(KL(J).EQ.0)GO TO 37
      S1=S1+CAI(I, J)*X(I1, J)
   37 CONTINUE
   38 FL(I)=ALOG(QI(I))+COI(I)+S1
      LX(I1)=1
      WN=BI(1)
      FLMAX=FL(1)
      DO 40 I=2, L
      IF(FL(I).GT.FLMAX)GO TO 39
      GO TO 40
   39 LX(I1)=I
      FLMAX=FL(I)
      WN=BI(I)
   40 CONTINUE
      IF(WN.EQ.R)RW=A
      IF(NI(L+1).EQ.0)GO TO 41
      IF(K.EQ.L+1)R=C
      IF(K.EQ.L+1)RW=C
   41 WRITE(2, 120)KI, R, WN, RW, FLMAX, LX(I1), (FL(I), I=1,
     -L)
  120 FORMAT(5X, I5, 9X, A1, 9X, A1, 9X, A1, F10.3, 5X, I5,
     -7(F10.3))
   42 CONTINUE
C     输出判别分类矩阵。
```

```
      WRITE(FM, '(A26, A21, I2, A11)')'(//10X, 12H判别分类矩阵/
     -4X, ', '17H判别分组/原始分组, ', L, 'I5, 5H小计)'
      WRITE(2, FM)(I, I=1, L)
      DO 44 K=1, L
      N1=LJ(K, 1)
      N2=LJ(K, 2)
      DO 43 J=1, L
      MB(K, J)=0
      DO 43 I=N1, N2
      IF(LX(I).EQ.J)MB(K, J)=MB(K, J)+1
   43 CONTINUE
   44 CONTINUE
      DO 45 K=1, L
      DO 45 J=1, L
      MB(K, L+1)=MB(K, L+1)+MB(K, J)
   45 MB(L+1, K)=MB(L+1, K)+MB(J, K)
      DO 46 K=1, L
   46 MB(L+1, L+1)=MB(L+1, L+1)+MB(L+1, K)
      DO 47 K=1, L
   47 WRITE(2, 121)K, (MB(K, J), J=1, L+1)
  121 FORMAT(11X, I1, 9X, 10I5)
      WRITE(2, 122)(MB(L+1, J), J=1, L+1)
  122 FORMAT(8X, '小计', 9X, 10I5)
 4000 WRITE(2, 100)
      RETURN
   48 WRITE(*, *)'您的数据文件有错！'
      END
```

(2)矩阵元素变换子程序

```
      SUBROUTINE VARY(A, B, M, KI, EP, KK)
      DIMENSION A(50, 50), B(50, 50)
      DO 6 I=1,M
      W=A(KI, KI)
      IF(W.LT.EP)GO TO 8
      IF(I.EQ.KI)GO TO 3
      DO 2 J=1, M
      IF(J.EQ.KI)GO TO 1
      B(I, J)=A(I, J)-A(I, KI)*A(KI, J)/W
      GO TO 2
    1 B(I, J)=-A(I, KI)/W
```

```
2 CONTINUE
  GO TO 6
3 DO 5 J=1, M
  IF(J.EQ.KI)GO TO 4
  B(I, J)=A(KI, J)/W
  GO TO 5
4 B(I, J)=1/W
5 CONTINUE
6 CONTINUE
  DO 7 I=1, M
  DO 7 J=1, M
7 A(I, J)=B(I, J)
  GO TO 9
8 KK=1
9 CONTINUE
  END
```

(3)生成F检验值的函数子程序

```
  FUNCTION FBIAO(N1, N2)
  INTEGER NEWN1(120), NEWN2(120)
  REAL BIK(34, 19)
  DATA NEWN1/1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 2*11, 3*12, 5*13,
 -4*14, 6*15, 10*16, 20*17, 60*18/
  DATA NEWN2/1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15,
  16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 10*31,
  20*32, 60*33/
  DATA MK/0/
  IF(MK.EQ.1)GO TO 1
  MK=1
  OPEN(5, FILE='F.DAT')
  READ(5, '(19F7.2)')((BIK(I, J), J=1, 19), I=1, 34)
  CLOSE(5)
1 IN1=19
  IN2=34
  IF(N1.LE.120)IN1=NEWN1(N1)
  IF(N2.LE.120)IN2=NEWN2(N2)
  FBIAO=BIK(IN2, IN1)
  END
```

六、逐步判别分析计算结果

判别组数= 3

变量个数= 4

样品总数=17

F检验方式= 1

第1组样品原始数据表

样品序号	变量1	变量2	变量3	变量4
1	6.000	−11.500	19.000	90.000
2	−4.000	−15.000	13.000	54.000
3	.000	−23.000	5.000	−35.000
4	−100.000	−21.400	7.000	−15.000
5	−5.000	−18.500	15.000	18.000
6	10.000	−18.000	14.000	50.000
7	−8.000	−14.000	16.000	56.000
变量平均值	−14.429	−17.343	12.714	31.143

第2组样品原始数据表

样品序号	变量1	变量2	变量3	变量4
1	90.200	−17.000	17.000	3.000
2	.000	−14.000	20.000	35.000
3	−100.000	−21.500	15.000	−40.000
4	13.000	−17.200	18.000	2.000
变量平均值	.800	−17.425	17.500	.000

第3组样品原始数据表

样品序号	变量1	变量2	变量3	变量4
1	−11.000	−18.500	25.000	−36.000
2	.500	−11.500	19.000	37.000
3	−10.000	−19.000	21.000	−42.000
4	20.000	−22.000	8.000	−20.000
5	.600	−13.000	26.000	21.000
6	−40.000	−20.000	22.000	−50.000
变量平均值	−6.650	−17.333	20.167	−15.000
变量总平均值	−8.100	−17.359	16.471	7.529

原始的组内离差矩阵

29042.4700	900.9814	419.2928	11076.6300
900.9814	215.4380	193.6976	1974.8430
419.2928	193.6976	373.2619	1435.2860
11076.6300	1974.8430	1435.2860	20894.8600

原始的总离差矩阵

29652.2800	898.1399	654.4999	9566.4990
898.1399	215.4612	193.5706	1976.0290
654.4999	193.5706	558.2353	283.7647

9566.4990　　1976.0290　　283.7647　　28070.2300

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

＊　　　　　　　　　　　　　　　＊

＊　引进变量的F检验值＝　2.00　＊

＊　　　　　　　　　　　　　　　＊

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

此次可以引进变量！计算结果如下：

变量序号	变量判别能力	是否已引进
1	.9794	(0)未引进
2	.9999	(0)未引进
3	.6686	(1)已引进
4	.7444	(0)未引进

即：未引进变量中第3号变量的判别能力最强，U=　　.6686

检验统计量 FU=　　3.4689大于FF1=　2.0000　　因此引进第3号变量！

＊＊

第1次变换后的组内离差矩阵

28571.4700	683.3969	−1.1233	9464.3430
683.3969	114.9221	−.5189	1230.0270
1.1233	.5189	.0027	3.8453
9464.3430	1230.0270	−3.8453	15375.8200

第1次变换后的总离差矩阵

28884.9100	671.1891	−1.1724	9233.8010
671.1891	148.3397	−.3468	1877.6330
1.1724	.3468	.0018	.5083
9233.8010	1877.6330	−.5083	27925.9900

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

＊　　　　　　　　　　　　　　　＊

＊　引进变量的F检验值＝　2.00　＊

＊　　　　　　　　　　　　　　　＊

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊

此次可以引进变量！计算结果如下：

变量序号	变量判别能力	是否已引进
1	.9891	(0)未引进
2	.7747	(0)未引进
3	.6686	(1)已引进
4	.5506	(1)已引进

即：未引进变量中第4号变量的判别能力最强，U=　　.5506

检验统计量FU=　　5.3055大于FF1=　2.0000　　因此引进第4号变量！

＊＊

第2次变换后的组内离差矩阵

22745.8500	−73.7265	1.2436	−.6155
−73.7265	16.5231	−.2113	−.0800
−1.2436	.2113	.0036	−.0003
.6155	.0800	−.0003	.0001

第2次变换后的总离差矩阵

25831.7300	50.3451	−1.0044	−.3307
50.3451	22.0952	−.3126	−.0672
1.0044	.3126	.0018	.0000
.3307	.0672	.0000	.0000

```
* * * * * * * * * * * * * * * * *
*                               *
*  剔出变量的F检验值＝   2.00   *
*                               *
* * * * * * * * * * * * * * * * *
```

此次计算结果表明，不能剔出变量！

即：已引进变量中第4号变量的判别能力最弱，U＝ .5506

检验统计量FU＝ 5.3055大于FF2＝ 2.0000 因此，不能剔出变量！

```
* * * * * * * * * * * * * * * * *
*                               *
*  引进变量的F检验值＝   2.00   *
*                               *
* * * * * * * * * * * * * * * * *
```

此次可以引进变量！计算结果如下：

变量序号	变量判别能力	是否已引进
1	.8805	(0)未引进
2	.7478	(1)已引进
3	.4946	(1)已引进
4	.5506	(1)已引进

即：未引进变量中第2号变量的判别能力最强，U＝ .7478

检验统计量FU＝ 2.0234大于FF1＝ 2.0000 因此引进第2号变量！

* *

第3次变换后的组内离差矩阵

22416.8800	4.4620	.3006	−.9725
−4.4620	.0605	−.0128	−.0048
−.3006	−.0128	.0063	.0008
.9725	−.0048	.0008	.0005

第3次变换后的总离差矩阵

25717.0200	−2.2786	−.2921	−.1775

2.2786	.0453	−.0141	−.0030
.2921	−.0141	.0062	.0009
.1775	−.0030	.0009	.0002

```
*****************
*               *
*  剔出变量的F检验值=  2.00  *
*               *
*****************
```

此次可以剔出变量！计算结果如下：

变量序号	变量判别能力	是否已引进
1	.8717	(0)未引进
2	.7478	(1)已引进
3	.9809	(0)未引进
4	.5316	(1)已引进

即：已引进变量中第3号变量的判别能力最弱，U= .9809

检验统计量FU= .1165小于FF2= 2.0000 因此剔出第3号变量！

```
****************************************
```

第4次变换后的组内离差矩阵

22431.1200	5.0682	−47.3934	−1.0091
−5.0682	.0347	2.0162	−.0033
−47.3934	−2.0162	157.6435	.1219
1.0091	−.0033	−.1219	.0004

第4次变换后的总的离差矩阵

25730.7300	−2.9427	46.9488	−.1336
2.9427	.0131	2.2735	−.0009
46.9488	−2.2735	160.7050	.1499
.1336	−.0009	−.1499	.0001

```
*****************
*               *
*  剔出变量的F检验值=  2.00  *
*               *
*****************
```

此次计算结果表明，不能剔出变量！

即：已引进变量中第2号变量的判别能力最弱，U= .3770

检验统计量FU= 10.7399大于FF2= 2.0000 因此，不能剔出变量！

```
*****************
*               *
*  引进变量的F检验值=  2.00  *
*               *
*****************
```

此次计算结果表明，不能引进变量！

即：未引进变量中第1号变量的判别能力最强U＝ .8718

检验统计量FU＝ .8826小于FF1＝ 2.0000 因此，不能引进变量！

各组判别函数值

分组序号	先验概率	COI项值	变量序号	CAI项值
1	.412	−100.388	1	.000
			2	−9.865
			3	.000
			4	.953

分组序号	先验概率	COI项值	变量序号	CAI项值
2	.235	−73.827	1	.000
			2	−8.474
			3	.000
			4	.801

分组序号	先验概率	COI项值	变量序号	CAI项值
3	.353	−61.667	1	.000
			2	−7.740
			3	.000
			4	.721

已知分组样品及待判样品的分类结果如下：

第1组样品的分类表

序号	原分组	判别分组	分组正误	最大f值	所属分组	1组f值	2组f值	3组f值
1	1	1	+	97.966	1	97.966	94.253	91.230
2	1	1	+	98.177	1	98.177	95.079	92.346
3	1	1	+	92.259	1	92.259	91.591	90.054
4	1	1	+	95.540	1	95.540	94.050	92.100
5	1	1	+	98.388	1	98.388	95.905	93.463
6	1	1	+	123.960	1	123.960	117.297	112.680
7	1	1	+	90.219	1	90.219	88.207	86.049

第2组样品的分类表

序号	原分组	判别分组	分组正误	最大f值	所属分组	1组f值	2组f值	3组f值
1	2	2	+	71.182	2	69.291	71.182	71.031
2	2	2	+	71.389	2	70.200	71.389	70.899
3	2	2	+	74.876	2	72.695	74.876	74.838
4	2	2	+	72.075	2	70.311	72.075	71.858

第3组样品的分类表

序号	原分组	判别分组	分组正误	最大f值	所属分组	1组f值	2组f值	3组f值
1	3	3	+	54.504	3	46.912	52.658	54.504
2	3	3	+	52.993	3	47.444	51.806	52.993

3	3	3	+	54.045	3	46.125	52.089	54.045
4	3	1	—	96.692	1	96.692	95.130	93.137
5	3	3	+	53.059	3	46.989	51.703	53.059
6	3	3	+	56.013	3	48.364	54.156	56.013

判别分类矩阵

判别分组／原始分组	1	2	3	小计
1	7	0	0	7
2	0	4	0	4
3	1	0	5	6
小计	8	4	5	17

七、F检验时的数据文件

如果KF＝0，要由函数子程序FBIAO计算产生FF1值进行F检验时，F.DAT数据文件需要存入当前工作盘.F.DAT数据文件请见上面的两组判别分析。

程序十二　R型因子分析

一、程序主要功能

通过因子分析可以把具有错综复杂关系的数量较多的因子（变量），归结为数量较少的几个综合因子，这些综合因子既可尽可能多地保留原有的因子信息，而且这些综合因子之间又是独立的。

因子分析有以下三方面的作用：

（1）压缩原始数据

因子分析可以把大量的原始数据在数量上大大的精简，以利于综合分析，从成因意义上讲，压缩了的数据在质量上提高了，而且可以保留原有数据的绝大多数成因信息。

（2）指示成因推理方向

因子分析能够把庞杂纷乱的原始数据，按成因上的联系进行归纳，整理，精炼与分类，理出几条比较客观的成因线索，因而可为研究人员提供推理方向，启发思考相应的成因结论。

（3）对叠加的地质过程进行分解

绝大多数情况下，地质现象都是多种地质过程的叠加产物，既有时间上的叠加，也有空间上的叠加。这些过程互相干扰，互相掩盖，使得每个地质过程的特征面貌不清，给成因研究带来了复杂性。而因子分析对解决这一问题却可以提供巧妙的途径。

二、程序符号说明

N ———————样品数（要求N小于500）；

M　——————变量数（要求M小于20）；

X(500,20)——原始数据矩阵，矩阵的行号为样品编号，矩阵的列号为变量编号；

QR(20,20)——相关系数矩阵；

D(20)————特征值；

V(20,20)　——特征向量；

B(20)————用雅可比法求特征值时的工作单元；

Z(20) ————用雅可比法求特征值时的工作单元；

Y(500,10) ——计算时的工作单元；

T(500,10) ——计算时的工作单元；

KK —————用雅可比法计算时的变换次数；

KA ——————令KA=1时，表示求特征值时也求特征向量；

令KA=0时，表示只求特征值；

SD ——————用户要求的特征值所占百分比；

EP ——————正交旋转的迭代精度。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D5-2.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
--------------------------------
N,M,SD
((X(I,J),J=1,M),I=1,N)
--------------------------------
```

例如，下面的数据文件（D5-2.DAT）就是一个供用户检测本程序的数据文件：

```
--------------------------------
20,7,0.90
11.835,0.48,14.36,25.21,25.21,0.81,0.98,
45.596,0.526,13.85,24.04,26.01,0.91,0.96,
3.525,0.086,24.4,49.3,11.30,6.82,0.85
3.681,0.37,13.57,26.12,26.00,0.82,1.01,
48.287,0.386,14.5,25.9,23.32,2.18,0.93,
17.956,0.28,9.75,17.05,37.2,0.464,0.98,
7.37,0.506,13.6,34.28,10.69,8.8,0.56,
4.223,0.34,3.8,7.1,88.2,1.11,0.97,
6.442,0.19,4.7,9.1,72.2,0.74,1.03,
16.234,0.39,3.1,5.4,121.5,0.42,1.00,
10.585,0.42,2.4,4.7,135.6,0.87,0.98,
23.535,0.23,2.6,4.6,141.8,0.31,1.02,
5.398,0.12,2.8,6.2,111.2,1.14,1.07
283.149,0.148,1.763,2.968,215.86,0.140,0.98,
316.604,0.317,1.453,2.432,263.41,0.249,0.98,
307.31,0.173,1.627,2.729,235.70,0.214,0.99,
322.515,0.312,1.382,2.320,282.21,0.024,1.00,
254.58,0.297,0.899,1.476,410.30,0.239,0.93,
304.092,0.283,0.789,1.357,438.36,0.193,1.01,
202.446,0.042,0.741,1.266,309.77,0.29,0.99
--------------------------------
```

四、计算结果输出

本程序输出文件的约定名为D5-2.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、R型因子分析主程序

```
      PROGRAM C9013
      COMMON X(500,20),QR(20,20),D(20),V(20,20),B(20),Z(20),
     -T(500,10),Y(500,10)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'R型因子分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名［约定名D5-2.DAT］：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D5-2.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名［约定名D5-2.WRI］：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D5-2.WRI'
      OPEN(2'FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入因子分析的原始数据；'
      READ(1,*,ERR=3)N,M,SD
      READ(1,*,ERR=3)((X(I,J),J=1,M),I=1,N)
      WRITE(*,*)'开始进行因子分析计算；请等待！'
      WRITE(2,*)'        **********************'
      WRITE(2,*)'        *                    *'
      WRITE(2,*)'        *    R 型因子分析计算结果    *'
      WRITE(2,*)'        *                    *'
      WRITE(2,*)'        **********************'
      WRITE(2,100)N,M,SD
  100 FORMAT(//5X,'样品数=',I3/5X,'变量数=',I3/
     -5X,'用户要求的特征值所占百分比=',F5.3)
      WRITE(2,'(//10X,A)')'原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      DO 2 I=1,N
    2 WRITE(2,102)I,(X(I,J),J=1,M)
  102 FORAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      CALL QRS(M,N,SD)
```

```
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？ [Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    3 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

(1) R 型因子分析子程序

```
      SUBROUTINE QRS(N,M,SD)
      COMMON X(500,20),QR(20,20),D(20),V(20,20),B(20),Z(20),
     -T(500,10),Y(500,10)
C     计算各个变量的平均值.
      DO 2 J=1,N
      B(J)=0.
      DO 1 I=1,M
    1 B(J)=B(J)+X(I,J)
    2 B(J)=B(J)/M
C     计算相关系数矩阵QR，因为QR是对称矩阵，在此只计算上三角部分的元素.
      DO 4 I=1,N
      DO 4 J=I,N
      S1=0.
      S2=0.
      S3=0.
      D0 3 K=1,M
      S1=S1+(X(K,I)-B(I))*(X(K,J)-B(J))
      S2=S2+(X(K,I)-B(I))*(X(K,I)-B(I))
    3 S3=S3+(X(K,J)-B(J))*(X(K,J)-B(J))
    4 QR(I,J)=S1/SQRT(S2*S3)
C     按对称关系，计算QR矩阵下三角部分的元素.
      DO 5 I=2,N
      I1=I-1
      DO 5 J=1,I1
    5 QR(I,J)=QR(J,I)
      WRITE(2,'(//10X,A)')'相关系数矩阵'
      DO 6 I=1,N
    6 WRITE(2,101)(QR(I,J),J=1,N)
  101 FORMAT(5X,10F12.4)
```

```
C     保留QR矩阵，将QR的值传给Y.
      DO 7 I=1,N
      DO 7 J=1,N
    7 Y(I,J)=QR(I,J)
      WRITE(*,*)'正在调用雅可比法子程序，计算特征值与特征向量.'
      CALL JACOBI(N,1,KK)
      WRITE(*,'(1X,A,I5)')'雅可比法计算完毕；旋转次数=',KK
      WRITE(2,'(//5X,A,I5)')'雅可比法旋转次数=',KK
      WRITE(2,'(//10X,A)')'相关系数矩阵的特征值'
      WRITE(2,101)(D(I),I=1,N)
      WRITE(2,'(//10X,A)')'相关系数矩阵的特征向量'
C     计算全部样品的特征值之和.
      S1=0.
      DO 8 I=1,N
      S1=S1+D(I)
    8 WRITE(2,101)(V(I,J),J=1,N)
C     按着由大到小的顺序，对特征值与特征向量进行排序.
      WRITE(2,'(//)')
      S2=0.
      DO 12 I=1,N
      DMAX=D(I)
      DO 11 J=I,N
      IF(D(J).GT.DMAX)GO TO 9
      GO TO 11
    9 DMAX=D(J)
      D(J)=D(I)
      D(I)=DMAX
      DO 10 K=1,N
      B(K)=V(K,J)
      V(K,J)=V(K,I)
   10 V(K,I)=B(K)
   11 CONTINUE
      S2=S2+D(I)
      S3=S2/S1
C     输出排序后的累计特征值.
      WRITE(2,102)I,S3
  102 FORMAT(5X,'前',I2,'个主因子累计特征值百分比=',F6.4)
      KK=I
C     当累计特征值大于用户要求的特征值所占百分比SD时，停止循环计算，转向标号
```

```
C      13.
       IF(S3.GT.SD)GO TO 13
    12 CONTINUE
    13 IF(KK.EQ.1)THEN
       WRITE(*,*)'因为用户要求的特征值所占比例太小，只能选出一个主因子.'
       WRITE(*,*)'请您放大SD值后，重新计算！'
         RETURN
       ENDIF
C      对特征向量矩阵进行规格化处理
       DO 14 J=1,KK
       Z(J)=SQRT(D(J)).
       DO 14 I=1,N
    14 V(I,J)=Z(J)*V(I,J)
       DO 16 I=1,N
       B(I)=0.
       DO 15 J=1,KK
    15 B(I)=B(I)+V(I,J)*V(I,J)
    16 Z(I)=SQRT(B(I))
       DO 17 I=1,N
       DO 17 J=1,KK
    17 V(I,J)=V(I,J)/Z(I)
C      以各个因子载荷值的平方方差之和达到最大，作为因子最大简化的标准
       VS=0.
       DO 19 J=1,KK
       S1=0.
       S2=0.
       DO 18 I=1,N
       S1=S1+(V(I,J)*V(I,J))**2
    18 S2=S2+V(I,J)*V(I,J)
    19 VS=VS+(N*S1-S2*S2)/(N*N)
       KS=0
       K1=KK-1
C      下面进行方差最大正交旋转.
  1000 VS1=VS
       KS=KS+1
       DO 26 K=1,K1
       K2=K+1
       DO 26 L=K2,KK
       AA=0.
```

```
      BB=0.
      CC=0.
      DD=0.
      DO 20 I=1,N
      AA=AA+V(I,K)*V(I,K)-V(I,L)*V(I,L)
      BB=BB+2*V(I,K)*V(I,L)
      CC=CC+(V(I,K)*V(I,K)-V(I,L)*V(I,L))**2-(2*V(I,K)*V(I,
     L))**2
   20 DD=DD+(V(I,K)*V(I,K)-V(I,L)*V(I,L))*(2*V(I,K)*V(I,L))
      DD=2*DD
      E=DD-2*AA*BB/N
      F=CC-(AA*AA-BB*BB)/N
      TAN4F=E/F
      IF(E.GE.0.0.AND.F.GE.0.0)F4=ATAN(TAN4F)
      IF(E.GE.0.0.AND.F.LT.0.0)F4=3.141592654+ATAN(TAN4F).
      IF(E.LT.0.0.AND.F.LT.0.0)F4=-3.141592654+ATAN(TAN4F)
      IF(E.LT.0.0.AND.F.GE.0.0)F4=ATAN(TAN4F)
      F1=F4/4
      S1=SIN(F1)
      S2=COS(F1)
      DO 21 I=1,KK
      DO 21  J=1,KK
      IF(I.EQ.J)T(I,J)=1.0
      IF(I.NE.J)T(I,J)=0.
      IF(I.EQ.K.AND.J.EQ.K)T(I,J)=S2
      IF(I.EQ.L.AND.J.EQ.L)T(I,J)=S2
      IF(I.EQ.K.AND.J.EQ.L)T(I,J)=-S1
      IF(I.EQ.L.AND.J.EQ.K)T(I,J)=S1
   21 CONTINUE
      DO 23 I=1,N
      DO 22 J=1,KK
      B(J)=V(I,J)
   22 V(I,J)=0.
      DO 23 J=1,KK
      DO 23 MM=1,KK
   23 V(I,J)=V(I,J)+B(MM)*T(MM,J)
      VS=0.
      DO 25 J=1,KK
      S1=0.
```

```
      S2=0.
      DO 24 I=1,N
      SI=S1+(V(I,J)*V(I,J))**2
   24 S2=S2+V(I,J)*V(I,J)
   25 VS=VS+(N*S1-S2*S2)/(N*N)
   26 CONTINUE
      IF(ABS(VS-VS1).GT.10E-7)GO TO 1000
C     计算相关矩阵的逆矩阵.
      DO 29 K=1,N
      W=QR(K,1)
      IF(W.LE.0.)GO TO 30
      DO 27 J=2,N
   27 QR(K,J-1)=QR(K,J)/W
      QR(K,N)=1/W
      DO 29 I=1,N
      IF(I.EQ.K)GO TO 29
      G=QR(I,1)
      DO 28 J=2,N
   28 QR(I,J-1)=QR(I,J)-QR(K,J-1)*G
      QR(I,N)=-QR(K,N)*G
   29 CONTINUE
      GO TO 31
   30 WRITE(*,*)'逆矩阵不存在，返回主程序，运行结束!'
      RETURN
   31 DO 32 I=1,N
      DO 32 J=1,KK
      Y(I,J)=0.
      DO 32 MM=1,N
   32 Y(I,J)=Y(I,J)+V(MM,J)*QR(MM,I)
      DO 33 I=1,M
      DO 33 J=1,KK
      T(I,J)=0.
      DO 33 MM=1,N
   33 T(I,J)=T(I,J)+Y(MM,J)*X(I,MM)
C     调用宽行打印因子图子程序，打印因子载荷图.
      DO 34 I=1,N
      DO 34 J=1,KK
   34 Y(I,J)=V(I,J)
      DO 35 K=1,KK
```

```
      LL=K+1
      DO 35 L=LL,KK
   35 CALL MAP(N,KK,K,L,0)
C     调用宽行打印因子图子程序，打印因子得分图.
      DO 36 I=1,M
      DO 36 J=1,KK
   36 Y(I,J)=T(I,J)
      DO 37 K=1,KK
      LL=K+1
      DO 37 L=LL,KK
   37 CALL MAP(M,KK,K,L,1)
      END
```

（2）雅可比法子程序

```
      SUBROUTINE JACOBI(N,KA,KK)
      COMMON X(500,20),QR(20,20),D(20),V(20,20),B(20),Z(20),
     -T(500,10),Y(500,10)
      IF(KA.EQ.0)GO TO 3
      DO 2 I=1,N
      DO 2 J=1,N
      IF(I.EQ.J)GO TO 1
      V(I,J)=0.
      GO TO 2
    1 V(I,J)=1.0
    2 CONTINUE
    3 DO 4 I=1,N
      D(I)=Y(I,I)
      B(I)=D(I)
    4 Z(I)=0.
      KK=0
      DO 25 K=1,100
      SM=0.
      N1=N-1
      DO 5 I=1,N1
      IP1=I+1
      DO 5 J=IP1,N
    5 SM=SM+ABS(Y(I,J))
      IF(SM)6,26,6
    6 IF(K-4)7,8,8
    7 TR=0.2*SM/(FLOAT(N)*FLOAT(N))
```

```
      GO TO 9
    8 TR=0
    9 DO 24 I=1,N1
      IP1=I+1
      DO 24 J=IP1,N
      G=100.0*ABS(Y(I,J))
      IF(K.GT.4.AND.ABS(D(I))+G.EQ.ABS(D(I)).AND.ABS(D(J))+G.EQ.
     -ABS(D(J)))GO TO 23
      IF(ABS(Y(I,J)).LE.TR)GO TO 24
      H=D(J)-D(I)
      IF(ABS(H)+G.EQ.ABS(H))GO TO 10
      TH=0.5*H/Y(I,J)
      TT=1.0/(ABS(TH)+SQRT(1.0+TH*TH))
      IF(TH.LT.0.)TT=-TT
      GO TO 11
   10 TT=Y(I,J)/H
   11 C=1.0/SQRT(1.0+TT*TT)
      S=TT*C
      H=TT*Y(I,J)
      Z(I)=Z(I)-H
      Z(J)=Z(J)+H
      D(I)=D(I)-H
      D(J)=D(J)+H
      Y(I,J)=0.
      I1=I-1
      IF(I1)14,14,12
   12 DO 13 L=1,I1
      G=Y(L,I)
      H=Y(L,J)
      Y(L,I)=C*G-S*H
   13 Y(L,J)=S*G+C*H
   14 J1=J-1
      IF(J1-IP1)17,15,15
   15 DO 16 L=IP1,J1
      G=Y(I,L)
      H=Y(L,J)
      Y(I,L)=C*G-S*H
   16 Y(L,J)=S*G+C*H
   17 JP1=J+1
```

```
      IF(N-JP1)20,18,18
   18 DO 19 L=JP1,N
      G=Y(I,L)
      H=Y(J,L)
      Y(I,L)=C*G-S*H
   19 Y(J,L)=S*G+C*H
   20 IF(KA.EQ.0)GO TO 22
      DO 21 L=1,N
      G=V(L,I)
      H=V(L,J)
      V(L,I)=C*G-S*H
   21 V(L,J)=S*G+C*H
   22 KK=KK+1
      GO TO 24
   23 Y(I,J)=0
   24 CONTINUE
      DO 25 I=1,N
      B(I)=B(I)+Z(I)
      D(I)=B(I)
   25 Z(I)=0.
   26 RETURN
      END
```

(3) 宽行打印因子图子程序

```
      SUBROUTINE MAP(N,KK,K,L,MM)
      COMMON X(500,20),QR(20,20),D(20),V(20,20),B(20),Z(20),
     -T(500,10),Y(500,10)
      DIMENSION XF(102),YF(52),W(101),IW(50)
      CHARACTER AA,BB,CC,DD,W
      DATA AA,BB,CC,DD/' ','*','I','+'/
      IF(MM.EQ.0)THEN
        WRITE(2,100)K,L
  100 FORMAT(//131('*')//5X,'变量序号',4X,'第',I2,'因子',4X,'第',
     -I2,'因子')
      ELSE
        WRITE(2,200)K,L
  200 FORMAT(//131('*')//5X,'样品序号',4X,'第',I2,'因子',4X,'第',
     -I2,'因子')
      ENDIF
      DO 1 I=1,N
```

```
  1 WRITE(2,101)I,Y(I,K),Y(I,L)
101 FORMAT(5X,I8,2F12.4)
    XMAX=Y(1,K)
    XMIN=Y(1,K)
    YMAX=Y(1,L)
    YMIN=Y(1,L)
    DO 2 I=1,N
    IF(Y(I,K).GT.XMAX)XMAX=Y(I,K)
    IF(Y(I,K).LT.XMIN)XMIN=Y(I,K)
    IF(Y(I,L).GT.YMAX)YMAX=Y(I,L)
    IF(Y(I,L).LT.YMIN)YMIN=Y(I,L)
  2 CONTINUE
    DX=(XMAX-XMIN)/100.
    DY=(YMAX-YMIN)/50.
    XMAX=XMAX+10*DX
    XMIN=XMIN-10*DX
    YMAX=YMAX+10*DY
    YMIN=YMIN-10*DY
    DX=(XMAX-XMIN)/100.
    DY=(YMAX-YMIN)/50.
    DO 3 I=1,102
  3 XF(I)=XMIN+DX*(I-1)
    DO 4 I=1,52
  4 YF(I)=YMAX-DY*(I-1)
    XF(1)=XF(1)-0.00001
    YF(1)=YF(1)+0.00001
    IF(MM.EQ.0)THEN
      WRITE(2,'(//50X,A,I2,A,I2,A)')'第',K,'与第',L,'因子载荷图'
    ELSE
      WRITE(2,'(//50X,A,I2,A,I2,A)')'第',K,'与第',L,'因子得分图'
    ENDIF
    WRITE(2,102)L
102 FORMAT(/8X,'第',I2,'因子')
    DO 19 I=1,50
    JJJ=0
    W(1)=CC
    DO 5 J=2,101
  5 W(J)=AA
    DO 9 J=1,N
```

```
      IF(Y(J,L).LT.YF(I).AND.Y(J,L).GE.YF(I+1))GO TO 6
      GO TO 9
    6 DO 8 K1=1,101
      IF(Y(J,K).GT.XF(K1).AND.Y(J,K).LE.XF(K1+1))GO TO 7
      GO TO 8
    7 W(K1)=BB
      JJJ=JJJ+1
      IW(JJJ)=J
      GO TO 9
    8 CONTINUE
    9 CONTINUE
      IF(JJJ.LE.1)GO TO 12
      JJ1=JJJ-1
      DO 11 J1=1,JJJ
      II=IW(J1)
      XXMIN=Y(II,K)
      DO 10 J2=J1,JJJ
      I1=IW(J2)
      IF(Y(I1,K).LT.XXMIN)THEN
        XXMIN=Y(I1,K)
      ENDIF
   10 CONTINUE
      DO 11 J2=J1,JJJ
      I1=IW(J2)
      IF(Y(I1,K)-XXMIN.LT.0.0000001)THEN
        I2=II
        IW(J1)=I1
        IW(J2)=I2
      ENDIF
   11 CONTINUE
   12 CONTINUE
      IF(MOD(I,5).EQ.1)GO TO 15
      IF(JJJ.EQ.0)GO TO 13
      IF(JJJ.GT.5)GO TO 14
      WRITE(2,106)(W(L1),L1=1,101),(IW(L1),L1=1,JJJ)
      GO TO 19
   13 WRITE(2,103)(W(L1),L1=1,101)
      GO TO 19
   14 WRITE(2,107)(W(L1),L1=1,101),(IW(L1),L1=1,5),JJJ
```

```
      GO TO 19
   15 IF(W(1).EQ.BB)GO TO 16
      W(1)=DD
   16 IF(JJJ.EQ.0)GO TO 17
      IF(JJJ.GT.5)GO TO 18
      WRITE(2,108)YE(I),(W(L1),L1=1,101),(IW(L1),L1=1,JJJ)
      GO TO 19
   17 WRITE(2,104)YF(I),(W(L1),L1=1,101)
      GO TO 19
   18 WRITE(2,109)YF(I),(W(L1),L1=1,101),(IW(L1),L1=1,5),JJJ
   19 CONTINUE
      WRITE(2,105)YF(51),(XF(L1),L1=1,101,10),K
  103 FORMAT(11X,101A1)
  104 FORMAT(1X,F10.2,101A1)
  105 FORMAT(1X,F10.2,1H+,10(10H----------+)/4X,11F10.2,1X,
     -'第','I2,因子')
  106 FORMAT(11X,101A1,5I3)
  107 FORMAT(11X,101A1,5I3,'(',I3,')')
  108 FORMAT(1X,F10.2,101A1,5I3)
  109 FORMAT(1X,F10.2,101A1,5I3,'(',I2,')')
      END
```

六、R 型因子分析计算结果（略）

程序十三　对 应 分 析

一、程序主要功能

对应分析是在R型因子分析与Q型因子分析基础上发展起来的一种多元统计分析方法，它可以由R型因子分析的结果，很容易地得到Q型因子分析的结果，这就克服了由于样品容量大而进行Q型分析所带来的计算上的困难。同时，对应分析把R型与Q型因子分析统一起来，把变量和样品同时反映在同一因子轴的图形上，这就便于进行地质解释与推断。在图形上邻近的变量点密切相关，具有成因上的联系，可能指示某一特定的地质作用或地质过程；类似地，在图形上邻进的一些样品点也密切相关，是同一地质过程的产物，可能同属一种类型。对于同属一种类型的样品点，可由邻进它们的变量点进行地质解释，同时通过样品在空间上的分布，也可以了解地质过程的空间关系。

二、程序符号说明

N——————样品数（要求N小于500）；

M——————变量数（要求M小于20）；

X(500,20)—原始数据矩阵，矩阵的行号为样品编号，矩阵的列号为变量编号；

XI(500)——每个样品的变量和（行和）；

XJ(20)——每个变量的样品和（列和）；

T ——————原始数据矩阵的元素总和；
Y(20,20) —乘积矩阵；
D(20) ———乘积矩阵的特征值；
V(20,20) —乘积矩阵的特征向量；
B(20) ———用雅可比法求特征值时的工作单元；
Z(20) ———用雅可比法求特征值时的工作单元；
XX(500,5)—计算时的工作单元；
KK ————用雅可比法计算时的变换次数；
KA ————令KA=1时，表示求特征值时也求特征向量，
令KA=0时，表示只求特征值；
SD ————用户要求的特征值所占百分比。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D6-1.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
N,M,SD
((X(I,J),J=1,M),I=1,N)
```

例如，下面的数据文件（D6-1.DAT）就是一个供用户检测本程序的数据文件：

```
15,4,0.95
11.835,14.36,25.21,25.21,
45.596,13.85,24.04,26.01,
3.525,24.4,49.3,11.30,
3.681,13.57,25.12,26.00,
48.287,14.5,25.9,23.32,
17.956,9.75,17.05,37.2,
7.37,13.6,34.28,10.69,
4.223,3.8,7.1,88.2,
6.442,4.7,9.1,73.2,
16.234,3.1,5.4,121.5,
10.585,2.4,4.7,135.6,
23.535,2.6,4.6,141.8,
5.398,2.8,6.2,111.2,
283.149,1.763,2.968,215.86,
202.446,0.741,1.266,309.77,
```

四、计算结果输出

本程序输出文件的约定名为D6-1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、对应分析主程序

```
      PROGRAM C9014
      COMMON X(500,20),XI(500),XJ(20),Y(20,20),D(20),V(20,20),B(20),
     —Z(20),XX(500,5)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,20(1H  )\)')
      WRITE(*,'(1X,A)')'对应分析'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名D6-1.DAT]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D6-1.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名[约定名D6-1.WRI]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D6-1.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入对应分析的原始数据！
      READ(1,*,ERR=3)N,M,SD
      READ(1,*,ERR=3)((X(I,J),J=1,M),I=1,N)
      WRITE(*,*)'开始进行对应分析计算；请等待！
      WRITE(2,*)'          **************'
      WRITE(2,*)'          *            *'
                          *            *
      WRITE(2,*)'          *  对应分析计算结果  *'
      WRITE(2,*)'          *            *'
                          *            *
      WRITE(2,*)'          **************'
      WRITE(2,100)N,M,SD
  100 FORMAT(//5X,'样品数=',I3/5X,'变量数=',I3/
     —5X,'用户要求的特征值所占百分比=',F5.2)
      WRITE(2,'(//10X,A)')'原始数据表'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      D0 2 I=1,N
    2 WRITE(2,102)I,(X(I,J),J=1,M)
  102 FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      CALL DYFX(N,M,SD)
```

```
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    3 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

(1) 对应分析子程序

```
      SUBROUTINE DYFX(N,M,SD)
      COMMON X(500,20),XI(500),XJ(20),Y(20,20),D(20),V(20,20),B(20),
     —Z(20),XX(500,5)
C     为求原始数据矩阵的元素总和T，首先令T等于零.
      T=0.
C     计算原始数据矩阵中每个样品的变量和XI，及全部元素总和T
      DO 2 I=1,N
      XI(I)=0.
      DO 1 J=1,M
    1 XI(I)=XI(I)+X(I,J)
    2 T=T+XI(I)
C     计算原始数据矩阵中每个变量的样品和XJ.
      DO 3 J=1,M
      XJ(J)=0.
      DO 3 I=1,N
    3 XJ(J)=XJ(J)+X(I,J)
      WRITE(2,'(//5X,A)')'原始数据矩阵及其行和,列和,总和如下：'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY),'行和'
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      DO 4 I=1,N
    4 WRITE(2,102)I,(X(I,J),J=1,M),XI(I)
  102 FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      WRITE(2,103)'列和',(XJ(J),J=1,M),T
  103 FORMAT(8X,A,1X,10(F10.3,1X)/(13X,10(F10.3,1X)))
C     对原始数据矩阵进行变换.
      DO 5 I=1,N
      DO 5 J=1,M
```

```
    5 X(I,J)=(X(I,J)-XI(I)*XJ(J)/T)/(SQRT(XI(I)*XJ(J)))
      WRITE(2,'(//10X,A)')'变换后的数矩矩阵'
      MXY=M
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
      DO 6 I=1,N
    6 WRITE(2,102)I,(X(I,J),J=1,M)
C     形成R型因子分析乘积矩阵.
      DO 7 I=1,M
      DO 7 J=1,M
      Y(I,J)=0.
      DO 7 K=1,N
    7 Y(I,J)=Y(I,J)+X(K,I)*X(K,J)
      WRITE(2,'(//10X,A)')'R型因子分析乘积矩阵'
      DO 8 I=1,M
    8 WRITE(2,104)(Y(I,J),J=1,M)
  104 FORMAT(5X,10F10.4)
      WRITE(*,*)'正在调用雅可比法子程序，计算特征值与特征向量 '
      CALL JACOBI(M,1,KK)
      WRITE(*,'(1X,A,I5)')'雅可比法计算完毕；旋转次数=',KK
      WRITE(2,'(//5X,A,I5)')'雅可比法旋转次数=',KK
      WRITE(2,'(//10X,A)')'乘积矩阵的特征值'
      WRITE(2,104)(D(I),I=1,M)
      WRITE(2,'(/10X,A)')'乘积矩阵的特征向量'
C     计算全部变量的特征值之和.
      SS=0.
      DO 9 I=1,M
      SS=SS+D(I)
    9 WRITE(2,104)(V(I,J),J=1,M)
C     按着由大到小的顺序，对特征值与特征向量进行排序
      WRITE(2,'(//)')
      S1=0.
      DO 13 I=1,M
      DMAX=D(I)
      DO 12 J=I,M
      IF(D(J).GT.DMAX)GO TO 10
      GO TO 12
   10 DMAX=D(J)
      D(J)=D(I)
```

```
      D(I)=DMAX
      DO 11 K=1,M
      XI(K)=V(K,J)
      V(K,J)=V(K,I)
   11 V(K,I)=XI(K)
   12 CONTINUE
      S1=S1+D(I)
      S2=S1/SS
C     输出排序后的累计特征值.
      WRITE(2,105)I,S2
  105 FORMAT(5X,'前',I2,'个主因子累计特征值百分比=',F6.4)
      KK=I
C     当累计特征值大于用户要求的特征值所占百分比SD时，停止循环计算，转向标号
C     14.
      IF(S2.GT.SD)GO TO 14
   13 CONTINUE
   14 IF(KK.EQ.1)THEN
      WRITE(*,*)'因为用户要求的特征值所占比例太小，只能选出一个主因子.'
      WRITE(*,*)'请您放大SD值后，重新计算！'
        RETURN
      ENDIF
C     如果主因子的个数大于5时，则令主因子数KK等于5.
      IF(KK.GT.5)KK=5
C     对特征向量矩阵进行规格化处理.
      DO 15 J=1,KK
      B(J)=SQRT(D(J))
      DO 15 I=1,M
   15 Y(I,J)=B(J)*V(I,J)
      DO 18 J=1,KK
      DO 16 I=1,N
      XI(I)=0.
      DO 16 K=1,M
   16 XI(I)=XI(I)+X(I,K)*V(K,J)
      S=0.
      DO 17 I=1,N
   17 S=S+XI(I)*XI(I)
      S=SQRT(S)
      DO 18 I=1,N
   18 XX(I,J)=XI(I)/S*B(J)
```

```
      DO 19 I=1,N
      DO 19 J=1,KK
   19 X(I,J)=XX(I,J)
C     调用宽行打图子程序，打印因子平面载荷图.
      DO 20 K=1,KK
      LL=K+1
      DO 20 L=LL,KK
   20 CALL MAP(N,M,KK,K,L)
      END
```

（2）雅可比法子程序

雅可比法子程序与R型因子分析相同.

（3）宽行打印因子图子程序

```
      SUBROUTINE MAP(N,M,KK,K,L)
      COMMON X(500,20),XI(500),XJ(20),Y(20,20),D(20),V(20,20),B(20),
     —Z(20),XX(500,5)
      DIMENSION XF(102),YF(52),W(101),IW(100)
      CHARACTER AA,BB,CC,DD,EE,W
      DATA AA,BB,CC,DD,EE/' ','*','I','+','O'/
      WRITE(2,100)K,L
  100 FORMAT(//131('*')//5X,'变量序号',4X,'第',I2,'因子',4X,'第',I2,
     —'因子')
      DO 1 I=1,M
    1 WRITE(2,101)I,Y(I,K),Y(I,L)
  101 FORMAT(10X,I3,2F12.4)
      WRITE(2,200)K,L
  200 FORMAT(//5X,'样品序号',4X,'第',I2,'因子',4X,'第',I2,'因子')
      DO 2 I=1,N
    2 WRITE(2,101)I,X(I,K),X(I,L)
      XMAX1=Y(1,K)
      XMIN1=Y(1,K)
      YMAX1=Y(1,L)
      YMIN1=Y(1,L)
      DO 3 I=1,M
      IF(Y(I,K).GT.XMAX1)XMAX1=Y(I,K)
      IF(Y(I,K).LT.XMIN1)XMIN1=Y(I,K)
      IF(Y(I,L).GT.YMAX1)YMAX1=Y(I,L)
      IF(Y(I,L).LT.YMIN1)YMIN1=Y(I,L)
    3 CONTINUE
      XMAX2=X(1,K)
```

```
      XMIN2=X(1,K)
      YMAX2=X(1,L)
      YMIN2=X(1,L)
      DO 4 I=1,N
      IF(X(I,K).GT.XMAX2)XMAX2=X(I,K)
      IF(X(I,K).LT.XMIN2)XMIN2=X(I,K)
      IF(X(I,L).GT.YMAX2)YMAX2=X(I,L)
      IF(X(I,L).LT.YMIN2)YMIN2=X(I,L)
    4 CONTINUE
      XMAX=AMAX1(XMAX1,XMAX2)
      XMIN=AMIN1(XMIN1,XMIN2)
      YMAX=AMAX1(YMAX1,YMAX2)
      YMIN=AMIN1(YMIN1,YMIN2)
      DX=(XMAX-XMIN)/100.
      DY=(YMAX-YMIN)/50.
      XMAX=XMAX+10*DX
      XMIN=XMIN-10*DX
      YMAX=YMAX+10*DY
      YMIN=YMIN-10*DY
      DX=(XMAX-XMIN)/100.
      DY=(YMAX-YMIN)/50.
      DO 5 I=1,102
    5 XF(I)=XMIN+DX*(I-1)
      DO 6 I=1,52
    6 YF(I)=YMAX-DY*(I-1)
      XF(1)=XF(1)-0.00001
      YF(1)=YF(1)+0.00001
      WRITE(2,'(//50X,A,I2,A,I2,A,)')'第',K,'与第',L,'因子载荷图
      WRITE(2,102)L
  102 FORMAT(/8X,'第',I2,'因子')
      DO 25 I=1,50
      JJJ=0
      W(1)=CC
      DO 7 J=2,101
    7 W(J)=AA
      DO 11 J=1,M
      IF(Y(J,L).LT.YF(I).AND.Y(J,L).GE.YF(I+1))GO TO 8
      GO TO 11
    8 DO 10 K1=1,101
```

```
      IF(Y(J,K).GT.XF(K1).AND.Y(J,K).LE.XF(K1+1))GO TO 9
      GO TO 10.
    9 W(K1)=EE
      GO TO 11
   10 CONTINUE
   11 CONTINUE
      DO 15 J=1,N
      IF(X(J,L).LT.YF(I).AND.X(J,L).GE.YF(I+1))GO TO 12
      GO TO 15
   12 DO 14 K1=1,101
      IF(X(J,K).GT.XF(K1).AND.X(J,K).LE.XF(K1+1))GO TO 13
      GO TO 14
   13 W(K1)=BB
      JJJ=JJJ+1
      IW(JJJ)=J
      GO TO 15
   14 CONTINUE
   15 CONTINUE
      IF(JJJ.LE.1)GO TO 18
      JJ1=JJJ-1
      DO 17 J1=1,JJJ
      II=IW(J1)
      XXMIN=X(II,K)
      DO 16 J2=J1,JJJ
      I1=IW(J2)
      IF(X(I1,K).LT.XXMIN)THEN
      XXMIN=X(I1,K)
      ENDIF
   16 CONTINUE
      DO 17 J2=J1,JJJ
      I1=IW(J2)
      IF(X(I1,K)-XXMIN.LT.0.0000001)THEN
      I2=II
      IW(J1)=I1
      IW(J2)=I2
      ENDIF
   17 CONTINUE
   18 CONTINUE
      IF(MOD(I,5).EQ.1)GO TO 21
```

```
      IF(JJJ.EQ.0)GO TO 19
      IF(JJJ.GT.5)GO TO 20
      WRITE(2,106)(W(L1),L1=1,101),(IW(L1),L1=1,JJJ)
      GO TO 25
   19 WRITE(2,103)(W(L1),L1=1,101)
      GO TO 25
   20 WRITE(2,107)(W(L1),L1=1,101),(IW(L1),L1=1,5),JJJ
      GO TO 25
   21 IF(W(1).EQ.BB.OR.W(1).EQ.EE)GO TO 22
      W(1)=DD
   22 IF(JJJ.EQ.0)GO TO 23
      IF(JJJ.GT.5)GO TO 24
      WRITE(2,108)YF(I),(W(L1),L1=1,101),(IW(L1),L1=1,JJJ)
      GO TO 25
   23 WRITE(2,104)YF(I),(W(L1),L1=1,101)
      GO TO 25
   24 WRITE(2,109)YF(I),(W(L1),L1=1,101),(IW(L1),L1=1,5),JJJ
   25 CONTINUE
      WRITE(2,105)YF(51),(XF(L1),L1=1,101,10),K
  103 FORMAT(11X,101A1)
  104 FORMAT(1X,F10.2,101A1)
  105 FORMAT(1X,F10.2,1H+,10(10H---------+)/4X,11F10.2,1X,
     -'第',I2,'因子')
  106 FORMAT(11X,101A1,5I3)
  107 FORMAT(11X,101A1,5I3,1H(,I2,1H))
  108 FORMAT(1X,F10.2,101A1,5I3)
  109 FORMAT(1X,F10.2,101A1,5I3,1H(,I2,1H))
      END
```

六、对应分析计算结果

样品数= 15

变量数= 4

用户要求的特征值所占百分比= .950

原始数据表

样品序号	变量 1	变量 2	变量 3	变量 4
1	11.835	14.360	25.210	25.210
2	45.596	13.850	24.040	26.010
3	3.525	24.400	49.300	11.300
4	3.681	13.670	25 120	26.000
5	48.287	14.600	25.900	23.320

6	17.956	9.750	17.050	37.200
7	7.370	13.600	34.280	10.690
8	4.223	3.800	7.100	88.200
9	6.442	4.700	9.100	73.200
10	16.234	3.100	5.400	121.500
11	10.585	2.400	4.700	135.600
12	23.535	2.600	4.600	141.800
13	5.398	2.800	6.200	111.200
14	283.149	1.763	2.968	215.860
15	202.446	.741	1.266	309.770

原始数据矩阵及其行和，列和，总和如下：

样品序号	变量 1	变量 2	变量 3	变量 4	行和
1	11.835	14.360	25.210	25.210	76.615
2	45.596	13.850	24.040	26.010	109.496
3	3.525	24.400	49.300	11.300	88.525
4	3.681	13.570	25.120	26.000	68.371
5	48.287	14.500	25.900	23.320	112.007
6	17.956	9.750	17.050	37.200	81.956
7	7.370	13.600	34.280	10.690	65.940
8	4.223	3.800	7.100	88.200	103.323
9	6.442	4.700	9.100	73.200	93.442
10	16.234	3.100	5.400	121.500	146.234
11	10.585	2.400	4.700	135.600	153.285
12	23.535	2.600	4.600	141.800	172.535
13	5.398	2.800	6.200	111.200	125.598
14	283.149	1.763	2.968	215.860	503.740
15	202.446	.741	1.266	309.770	514.223
列和	690.262	125.934	242.234	1356.860	2415.290

变换后的数矩阵

样品序号	变量 1	变量 2	变量 3	变量 4
1	−.044	.106	.129	−.055
2	.052	.069	.080	−.092
3	−.088	.187	.276	−.111
4	−.073	.108	.142	−.041
5	.059	.073	.089	−.102
6	−.023	.054	.063	−.027
7	−.054	.112	.219	−.088
8	−.095	−.014	−.021	.081
9	−.080	−.002	−.002	.058

10	−.080	−.033	−.049	.088
11	−.102	−.040	−.055	.109
12	−.075	−.043	−.062	.093
13	−.104	−.030	−.037	.098
14	.236	−.097	−.136	−.081
15	.093	−.102	−.143	.025

R 型因子分析乘积矩阵

.1375	−.0463	−.0694	−.0546
−.0463	.1090	.1562	−.0662
−.0694	.1562	.2291	−.0949
−.0546	−.0662	−.0949	.0992

雅可比法旋转次数= 23

乘积矩阵的特征值

.0016	.0000	.3969	.1763

乘积矩阵的特征向量

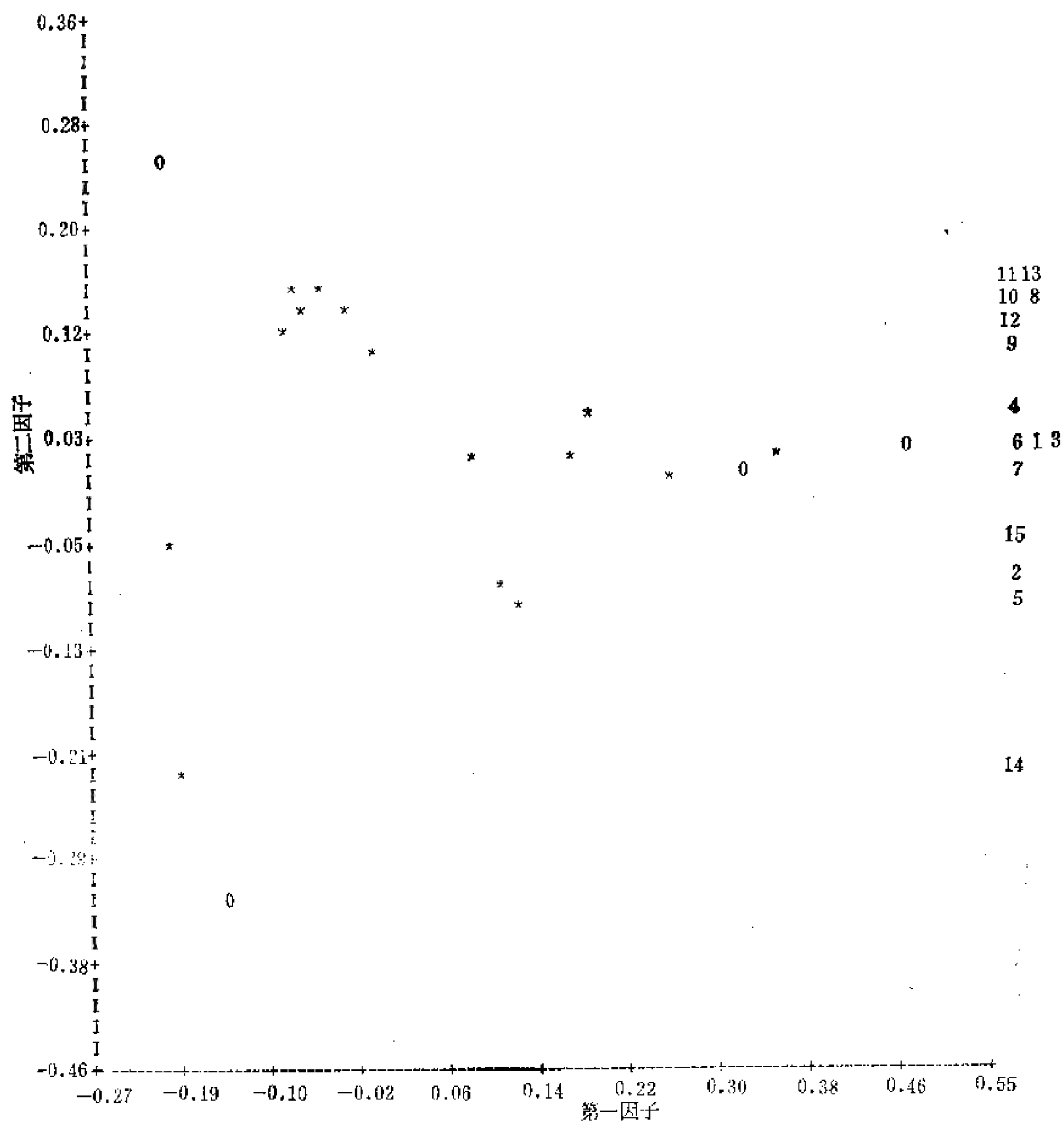
.0113	.5346	−.2295	−.8133
−.8222	.2283	.5213	−.0084
.5691	.3167	.7588	.0019
.0020	.7495	−.3158	.5818

前1个主因子累计特征值百分比= .6905

前2个主因子累计特征值百分比= .9971

变量序号	第 1因子	第 2因子
1	−.1446	−.3415
2	.3284	−.0035
3	.4781	.0008
4	−.1939	.2443

样品序号	第 1因子	第 2因子
1	.1801	.0028
2	.1141	−.0963
3	.3624	.0061
4	.1935	.0350
5	.1242	−.1072
6	.0893	.0029
7	.2644	−.0080
8	−.0266	.1240
9	−.0022	.0987
10	−.0642	.1170
11	−.0738	.1464
12	−.0819	.1149



附图9 对应分析的第 1与第2 因子载荷图

13	—.0507	.1417
14	—.1826	—.2386
15	—.1908	—.0606

程序十四 非线性映射

一、程序主要功能

非线性映射是一种几何图象降维方法，即通过非线性变换后，把高维空间中样品点形成的几何图像，变换成低维(二维或三维)空间中的图像。要求变换后仍能近似的保持原像的几何关系，这种方法直观形象，能使在低维空间上看到高维空间中样品点之间相互关系的近似图像。

二、程序符号说明

N——————样品数（要求N小于100）；

M——————变量数（要求M小于20）；

L——————低维空间的维数，为便于在平面上显示可令：L=2.

X(100,20)—原始数据矩阵，矩阵的行号为样品编号，矩阵的列号为变量编号；

Y(100,2)—二维空间上N个样品点的座标（要求给定初始值）；

D(100,100)—距离系数矩阵，主对角线原素为零；

上三角部分存高维空间中N个样品间的距离系数，

下三角部分存低维空间中N个样品间的距离系数；

FN —————高维空间中N个样品间的距离总和；

FM —————魔力因子，即为控制迭代收敛速度的经验系数，

一般可令FM=0.3—0.4；

KK —————迭代次数，本程序限定KK超过100次时，计算结束；

XK —————从高维空间变换到低维空间的约束条件，也称误差函数，XK表示低维空间中N个样品点的构形与高维空间中N个样品点的构形之间的拟合程度。

EP —————迭代计算要求的精度。

三、数据文件格式

使用本程序时，用户必须按下列格式组成一个供计算使用的数据文件。本程序数据文件的约定名为D7-1.DAT，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。数据文件要按自由格式将输入数据按行排列如下：

```
N,M,L,FM
((X(I,J),J=1,M),I=1,N)
```

例如，下面的数据文件（D7-1.DAT）就是一个供用户检测本程序的数据文件：

```
9,6,2,0,4
0.000,0.000,0.000,0.000,0.000,0.000
0.252,0.462,-0.308,0.007,-0.441,0.143.
0.610,0.963,-0.650,-0.278,-0.444,0.056,
0.864,1.212,-0.827,-0.686,-0.008,-0.210,
0.866,1.061,-0.735,-0.980,0.612,-0.500,
0.614,0.599,-0.427,-0.986,1.053,-0.643,
0.256,0.097,-0.085,-0.702,1.056,-0.556,
0.002,-0.151,0.092,-0.293,0.620,-0.290,
0.443,0.530,-0.367,0.490,0.306,-0.260
1.000,1.000,
0.000,1.000,
-1.000,1.000,
```

```
-1.000,0.000,
-1.000,-1.000,
0.000,-1.000,
1.000,-1.000,
1.000,0.000,
0.000,0.000,
```

四、计算结果输出

本程序输出文件的约定名为D7-1.WRI，如果使用其他名称，要在程序执行时，由键盘录入指定的文件名。

五、非线性映射主程序

```
      PROGRAM C9015
      COMMON X(100,20),D(100,100),Y(100,2))Y1(100,20)
      CHARACTER FILENA*20,NOYES
    1 WRITE(*,'(1X,30(1H  )\)')
      WRITE(*,'(1X,A)')'非线性映射'
      WRITE(*,'(1X,A\)')'请输入您的数据文件名[约定名D7-1.DAT]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D7-1.DAT'
      OPEN(1,FILE=FILENA)
      WRITE(*,'(1X,A\)')'请输入您的输出文件名[约定名D7-1.WRI]：'
      READ(*,'(A)')FILENA
      IF(FILENA.EQ.' ')FILENA='D7-1.WRI'
      OPEN(2,FILE=FILENA,STATUS='NEW')
      WRITE(*,*)'开始读入非线性映射的原始数据：'
      READ(1,*,ERR=4)N,M,L,FM
      READ(1,*,ERR=4)((X(I,J),J=1,M),I=1,N)
      READ(1,*,ERR=4)((Y(I,J),J=1,L),I=1,N)
      WRITE(*,*)'开始进行非线性映射计算，请等待！'
      WRITE(2,*)'          ********************'
      WRITE(2,*)'          *                  *'
      WRITE(2,*)'          *  非线性映射计算结果  *'
      WRITE(2,*)'          *                  *'
      WRITE(2,*)'          ********************'
      WRITE(2,100)N,M,L,FM
  100 FORMAT(//5X,'样品数=',I3/5X,'变量数=',I3/5X,'低维空间的维数=',
     -I1/5X,'魔力因子=',F5.3)
      WRITE(2,'(//9X,A)')'高维空间的原始图像数据表'
      MXY=M
```

```
      IF(M.GT.10)MXY=10
      WRITE(2,101)'样品序号',('变量',J,J=1,MXY)
  101 FORMAT(/5X,A,10(4X,A,I2,1X))
      DO 2 I=1,N
    2 WRITE(2,102)I,(X(I,J),J=1,M)
  102 FORMAT(9X,I3,1X,10(F10.3,1X)/13X,10(F10.3,1X)))
      WRITE(2,'(//5X,A)')'二维空间的初始给定数据表'
      WRITE(2,101)'样品序号',('变量',J,J=1,L)
      DO 3 I=1,N
    3 WRITE(2,102)I,(Y(I,J),J=1,L)
      CALL FXXYS(N,M,L,FM)
      CLOSE(1)
      CLOSE(2)
      WRITE(*,'(1X,A\)')'程序运行完闭！还继续进行计算吗？[Y/N]：'
      READ(*,'(A)')NOYES
      IF(NOYES.EQ.'Y'.OR.NOYES.EQ.'y')GO TO 1
      STOP
    4 WRITE(*,*)'您的数据文件有错！'
      STOP
      END
```

(1) 非线性映射子程序

```
      SUBROUTINE FXXYS(N,M,L,FM)
      COMMON X(100,20),D(100,100),Y(100,2),Y1(100,20)
C     给迭代次数KK，高维空间全部样品的距离总和FN赋初值.
      KK=0
      FN=0.
C     下面的EP为迭代计算要求的精度.
      EP=1E-6
C     计算高维空间样品之间的距离系数，计算结果存于距离系数矩阵D的上三角部 分；
C     同时，计算高维空间全部样品的距离总和FN.
      DO 2 I=1,N
      DO 2 J=I,N
      D(I,J)=0.
      DO 1 K=1,M
    1 D(I,J)=D(I,J)+(X(I,K)-X(J,K))**2
      D(I,J)=SQRT(D(I,J))
    2 FN=FN+D(I,J)
      WRITE(2,'(//5X,A,F12.4)')'高维空间全部样品的距离总和=',FN
C     未达到迭代计算要求的精度或者迭代次数不大于100次时，继续进行迭代计算.
```

```
1000 DO 4 I=2,N
C       计算二维空间样品之间的距离系数，计算结果存于距离系数矩阵D的下三角部 分；
        I1=I-1
        DO 4 J=1,I1
        D(I,J)=0.
        DO 3 K=1,L
      3 D(I,J)=D(I,J)+(Y(I,K)-Y(J,K))**2
      4 D(I,J)=SQRT(D(I,J))
        IF(KK.EQ.0)THEN
        WRITE(2,'(//10X,A)')'距离系数矩陈'
        DO 5 I=1,N
      5 WRITE(2,101)(D(I,J),J=1,N)
    101 FORMAT(5X,10F12.4)
        WRITE(2,'(/5X,A)')'注：矩阵上三角部分为高维空间样品之间的距离系数'
        WRITE(2,'(5X,A//)')'        下三角部分为二维空间样品之间的距离系数'
        ENDIF
C       给空间变换约束条件XK(即误差函数)赋初值。
        XK=0
        N1=N-1
C       计算误差函数的初始值.
        DO 6 I=1,N1
        I1=I+1
        DO 6 J=I1,N
      6 XK=XK+((D(I,J)-D(J,I))**2)/D(I,J)
        XK=XK/FN
        WRITE(2,104)KK,XK
    104 FORMAT(5X,'第',I3,'次迭代后的误差函数值=',F12.6)
C       如果迭代次数KK>100时，则停止迭代，转入调用打印二维空间样品点图像子程序.
        IF(KK.GT.100)GO TO 12
C       如果误差函数XK<EP时，则停止迭代，转入调用打印二维空间样品点图像子程序.
        IF(XK.LT.EP)GO TO 12
C       进行迭代计算，使误差函数值逐渐达到极小.
        DO 10 I=1,N
        DO 10 J=1,L
        S1=0.
        S2=0.
        DO 9 K=1,N
        IF(K.EQ.I)GO TO 9
        IF(K.LT.I)GO TO 7
```

```
      ID=I
      KD=K
      DO TO 8
    7 ID=K
      KD=I
    8 D1=D(ID,KD)-D(KD,ID)
      D2=D(ID,KD)*D(KD,ID)
      D3=Y(I,J)-Y(K,J)
      S1=S1+(D1/D2)*D3
      S2=S2+(D1-(D3**2/D(KD,ID))*(1+D1/D(KD,ID)))/D2
    9 CONTINUE
   10 Y1(I,J)=Y(I,J)-FM*S1/S2
      DO 11 I=1,N
      DO 11 J=1,L
   11 Y(I,J)=Y1(I,J)
C     迭代次数累加一次
      KK=KK+1
      GO TO 1000
C     调用打印二维空间样品点图像子程序
   12 CALL MAP(N,L)
      END
```

(2) 打印二维空间样品点图像子程序

```
      SUBROUTINE MAP(N,L)
      COMMON X(100,20),D(100,100),Y(100,2),Y1(100,20)
      DIMENSION XX(102),YY(52),W(101),IW(100)
      CHARACTER AA,BB,CC,DD,W
      DATA AA,BB,CC,DD/' ','*','I','+'/
      WRITE(2,100)
  100 FORMAT(//131('*')//7X,'非线性映射变换后的图像数据表'/5X,
     -'样品序号',6X,'变量1',6X,'变量2')
      DO 1 I=1,N
    1 WRITE(2,101)I,(Y(I,J),J=1,L)
  101 FORMAT(10X,I3,2F12.4)
      XMAX=Y(1,1)
      XMIN=Y(1,1)
      YMAX=Y(1,2)
      YMIN=Y(1,2)
      DO 2 I=1,N
      IF(Y(I,1).GT.XMAX)XMAX=Y(I,1)
```

```
      IF(Y(I,1).LT.XMIN)XMIN=Y(I,1)
      IF(Y(I,2).GT.YMAX)YMAX=Y(I,2)
      IF(Y(I,2).LT.YMIN)YMIN=Y(I,2)
    2 CONTINUE
      DX=(XMAN-XMIN)/100.
      DY=(YMAX-YMIN)/50.
      XMAX=XMAX+15*DX
      XMIN=XMIN-15*DX
      YMAX=YMAX+15*DY
      YMIN=YMIN-15*DY
      DX=(XMAX-XMIN)/100.
      DY=(YMAX-YMIN)/50.
      DO 3 I=1,101
    3 XX(I)=XMIN+DX*(I-1)
      DO 4 I=1,51
    4 YY(I)=YMAX-DY*(I-1)
      XX(1)=XX(1)-0.00001
      YY(1)=YY(1)+0.00001
      WRITE(2,102)
  102 FORMAT(//50X,'样品点的二维空间图像')
      WRITE(2,103)
  103 FORMAT(/6X,'变量2')
      DO 19 I=1,50
      JJJ=0
      W(1)=CC
      DO 5 J=2,101
    5 W(J)=AA
      DO 9 J=1,N
      IF(Y(J,2).LT.YY(I).AND.Y(J,2).GE.YY(I+1))GO TO 6
      GO TO 9
    6 DO 8 K=1,101
      IF(Y(J,1).GT.XX(K).AND.Y(J,1).LE.XX(K+1))GO TO 7
      GO TO 8
    7 W(K)=BB
      JJJ=JJJ+1
      IW(JJJ)=J
      GO TO 9
    8 CONTINUE
    9 CONTINUE
```

```
      IF(JJJ.LE.1)GO TO 12
      JJ1=JJJ-1
      DO 11 J1=1,JJJ
      II=IW(J1)
      XXMIN=Y(II,1)
      DO 10 J2=J1,JJJ
      I1=IW(J2)
      IF(Y(I1,1).LT.XXMIN)THEN
      XXMIN=Y(I1,1)
      ENDIF
   10 CONTINUE
      DO 11 J2=J1,JJJ
      I1=IW(J2)
      IF(Y(I1,1)-XXMIN.LT.0.0000001)THEN
      I2=II
      IW(J1)=I1
      IW(J2)=I2
      ENDIF
   11 CONTINUE
   12 CONTINUE
      IF(MOD(I,5).EQ.1)GO TO 15
      IF(JJJ.EQ.0)GO TO 13
      IF(JJJ.GT.5)GO TO 14
      WRITE(2,107)(W(K),K=1,101),(IW(K),K=1,JJJ)
      GO TO 19
   13 WRITE(2,104)(W(K),K=1,101)
      GO TO 19
   14 WRITE(2,108)(W(K),K=1,101),(IW(K),K=1,5),JJJ
      GO TO 19
   15 IF(W(1).EQ.BB)GO TO 16
      W(1)=DD
   16 IF(JJJ.EQ.0)GO TO 17
      IF(JJJ.GT.5)GO TO 18
      WRITE(2,109)YY(I),(W(K),K=1,101),(IW(K),K=1,JJJ)
      GO TO 19
   17 WRITE(2,105)YY(I),(W(K),K=1,101)
      GO TO 19
   18 WRITE(2,110)YY(I),(W(K),K=1,101),(IW(K),K=1,5),JJJ
   19 CONTINUE
```

```
      WRITE(2,106)YY(51),(XX(K),K=1,101,10)
104   FORMAT(11X,101A1)
105   FORMAT(1X,F10.2,101A1)
106   FORMAT(1X,F10.2,1H+,10(10H----------+),1X,'变量1'/4X,
      11F10.2)
107   FORMAT(11X,101A1,5I3)
108   FORMAT(11X,101A1,5I3,1H(,I2,1H))
109   FORMAT(1X,F10.2,101A1,5I3)
110   FORMAT(1X,F10.2,101A1,5I3,1H(,I2,1H))
      END
```

六、非线性映射计算结果

样品数＝ 9

变量数＝ 6

低维空间的维数＝2

魔力因子＝ .400

高维空间的原始图像数据表

样品序号	变量 1	变量 2	变量 3	变量 4	变量 5	变量 6
1	.000	.000	.000	.000	.000	.000
2	.252	.462	-.308	.007	-.441	-.143
3	.610	.963	-.650	-.278	-.444	-.056
4	.864	1.212	-.827	-.686	-.008	-.210
5	.866	1.061	-.735	-.980	.612	.500
6	.614	.599	-.427	-.986	1.053	-.643
7	.256	.097	-.085	-.702	1.056	-.556
8	.002	-.151	.092	-.293	.620	-.290
9	.443	.530	-.367	-.490	.306	-.250

二维空间的初始给定数据表

样品序号	变量 1	变量 2
1	1.000	1.000
2	.000	1.000
3	-1.000	1.000
4	-1.000	.000
5	-1.000	-1.000
6	.000	-1.000
7	1.000	-1.000
8	1.000	.000
9	.000	.000

高维空间全部样品的距离总和＝ 48.2193

距离系数矩阵

.0000	.7660	1.4140	1.8477	2.0002	1.8473	1.4139	.7653	1.0040
1.0000	.0000	.7648	1.4138	1.8479	2.0000	1.8480	1.4147	1.0020
2.0000	1.0000	.0000	.7650	1.4140	1.8475	1.9997	1.8475	.9983
2.2361	1.4142	1.0000	.0000	.7656	1.4147	1.8481	2.0002	.9963
2.8284	2.2361	2.0000	1.0000	.0000	.7660	1.4147	1.8481	.9963
2.2361	2.0000	2.2361	1.4142	1.0000	.0000	.7651	1.4138	.9980
2.0000	2.2361	2.8284	2.2361	2.0000	1.0000	.0000	.7652	1.0015
1.0000	1.4142	2.2361	2.0000	2.2361	1.4142	1.0000	.0000	1.0040
1.4142	1.0000	1.4142	1.0000	1.4142	1.0000	1.4142	1.0000	.0000

注：矩阵上三角部分为高维空间样品之间的距离系数

下三角部分为二维空间样品之间的距离系数

第 0次迭代后的误差函数值＝ .074070

第 1次迭代后的误差函数值＝ .015050

第 2次迭代后的误差函数值＝ .003406

第 3次迭代后的误差函数值＝ .000979

第 4次迭代后的误差函数值＝ .000324

第 5次迭代后的误差函数值＝ .000113

第 6次迭代后的误差函数值＝ .000040

第 7次迭代后的误差函数值＝ .000014

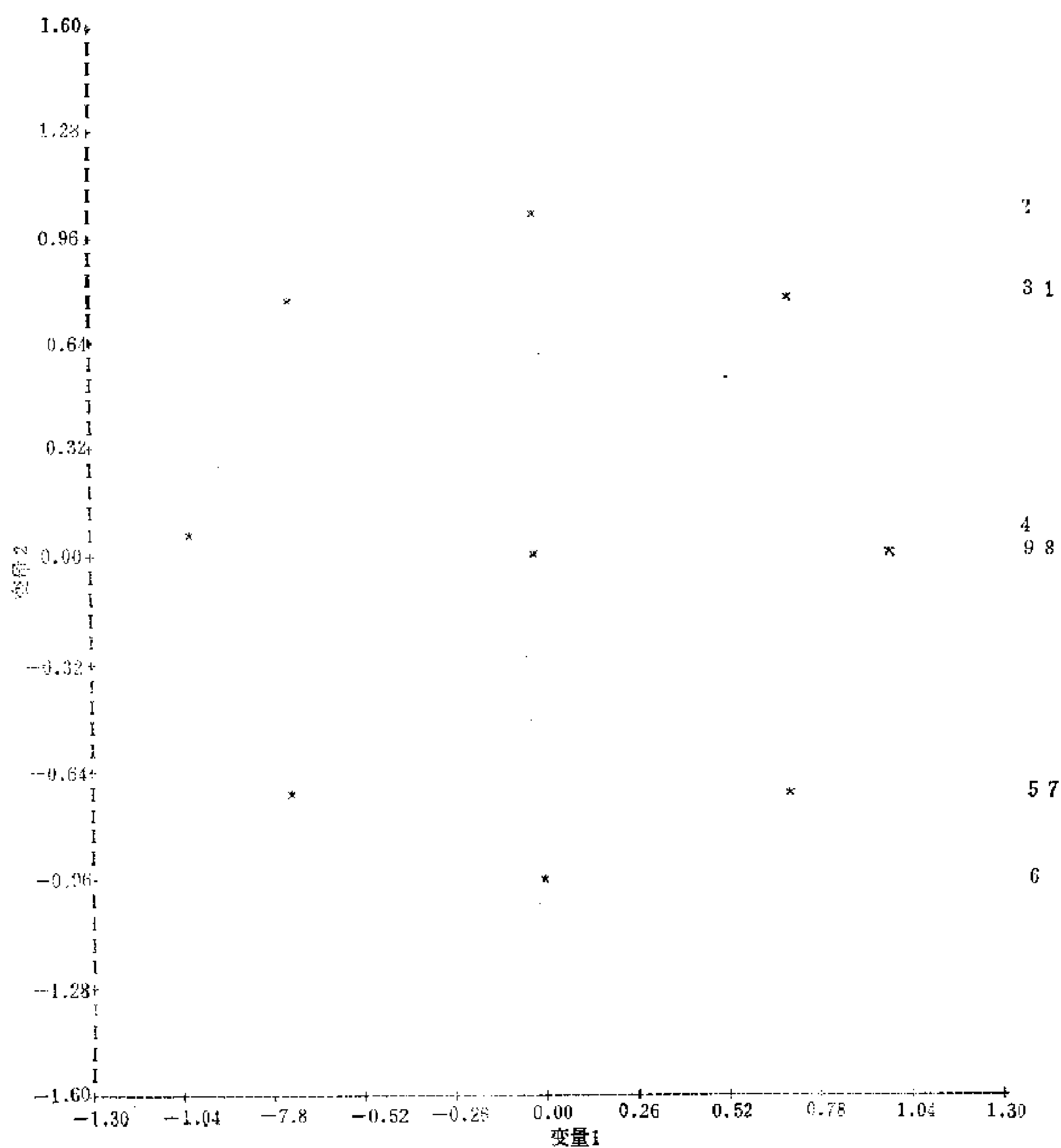
第 8次迭代后的误差函数值＝ .000005

第 9次迭代后的误差函数值＝ .000002

第 10次迭代后的误差函数值＝ .000001

非线性映射变换后的图像数据表

样品序号	变量 1	变量 2
1	.7084	.7079
2	−.0001	.9992
3	−.7072	.7078
4	−.[illegible]985	.0006
5	−.7077	−.7076
6	.0009	−.9985
7	.7084	−.7076
8	.9994	.0001
9	−.0034	−.0016



附图10　样品点的二维空间图像